REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

...o average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and ... Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions ... Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to ... Budget, Washington, DC 20503

2. REPORT DATE
15 May 1991

3. REPORT TYPE AND DATES COVERED
Final 23 Aug 89-30 Apr 91

4. TITLE AND SUBTITLE
QUANTITATIVE STRUCTURE-RETENTION RELATIONSHIPS FOR POLYCHLORINATED DIBENZODIOXINS AND POLYCHLORINATED DIBENZOFURANS

5. FUNDING NUMBERS

6. AUTHOR(S)
Mark D. Needham, CPT, U.S. Army

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
USASD, Fort Benjamin Harrison, IN

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Education Branch, DA-OPE-R
Perscom, Alexandria, VA

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

This thesis describes research involving Quantitative Structure-Retention Relationships (QSRR). This research utilizes applied mathematics, multivariate statistics and computational techniques for determining models of retention behavior. The methodology is based upon the fact that there is a relationship between a compound's molecular structure and its chromatographic retention behavior. Linear models are created relating the observed retention data to a set of descriptors which numerically encode structural information. The regression models are validated to ensure their credibility. The first study generated regression equations modeling gas chromatographic retention behavior of polychlorinated dibenzodioxins for non-polar, moderately polar, and polar stationary phases. Extremely accurate predictive models were validated and predictions made where no experimental data existed. The second study involved the complete set of polychlorinated dibenzofurans. Linear regression models were developed relating retention behavior on five chromatographic columns to descriptors which encoded the structural information. These models were then used to predict unknown retention data.

14. SUBJECT TERMS
Dibenzodioxins, dibenzofurans, gas chromatographic retention, prediction of retention data, QSRR.

15. NUMBER OF PAGES
156

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT UNCLASSIFIED	18. SECURITY CLASSIFICATION OF THIS PAGE UNCLASSIFIED	19. SECURITY CLASSIFICATION OF ABSTRACT UNCLASSIFIED	20. LIMITATION OF ABSTRACT UL

The Pennsylvania State University

The Graduate School

Department of Chemistry

QUANTITATIVE STRUCTURE RETENTION RELATIONSHIPS OF POLYCHLORINATED DIBENZODIOXINS AND DIBENZOFURANS

A Thesis in

Chemistry

by

Mark D. Needham

Submitted in Partial Fullfillment
of the Requirements
for the Degree of

Master of Science

August 1991

Mark D. Needham

Mark D. Needham

We approve the thesis of Mark D. Needham.

Date of Signature

Peter C Jurs — April 30, 1991

Peter C. Jurs
Professor of Chemistry
Thesis Adviser

P S Weiss — April 30, 1991

Paul S. Weiss
Assistant Professor of Chemistry

C. Robert Matthews — April 30, 1991

C. Robert Matthews
Professor of Chemistry
Eberly Family Professor
of Biotechnology

Barbara J Garrison — April 30, 1991

Barbara J. Garrison
Professor of Chemistry
Head of the Department of Chemistry

ABSTRACT

This thesis describes research involving Quantitative Structure-Retention Relationships (QSRR). This type of research utilizes applied mathematics, multivariate statistics and computational techniques for determining models of retention behavior.

The methodology discussed here is based upon the fact that there is a relationship between a compound's molecular structure and its chromatographic retention behavior. Linear models are created relating the observed retention data to a set of descriptors which numerically encode structural information. The regression models are statistically validated to ensure their credibility. Statistical transformations, used to improve a model's predictability, are also discussed.

The first study generated regression equations modeling gas chromatographic retention behavior of polychlorinated dibenzodioxins for non-polar, moderately polar, and polar stationary phases. Extremely accurate predictive models using topological, electronic, geometrical and atom-based descriptors were validated and predictions made for the isomers where no experimental data existed.

The second study involved the complete set of polychlorinated dibenzofurans. Linear regression models were developed to relate retention behavior on five chromatographic columns to a set of descriptors which encoded the structural information. These high quality models were then used to predict unknown retention data.

This thesis demonstrates the usefulness of computer assisted QSRR. The techniques described here are valuable tools for predicting unknown retention data.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank Professor Peter C. Jurs, my research advisor, for his support during my brief graduate career. His professional guidance is truly appreciated. I would also like to thank some of the other members of the research group whose guidance and assistance were an instrumental part of my research. In particular, I would like to acknowledge Dave Stanton and Paul Sutton, both of whom gave me a great start on these projects. Additionally, I would like to acknowledge Larry Anker for his ever-present assistance, understanding, and keen sense of humor.

I sincerely appreciate the enjoyable experience I have had in graduate school due to efforts of Jon Ball, Leanne Egolf, Rich Lawson, Marty Ranc, Steve Dixon, Tom Woloszyn, Sue Moyer, Wally Murray, Takeshi Sakaki, and Jan Main.

This research was supported in part by the United States Army, which paid my full salary and tuition for my graduate stay. The Sun 4/110 workstation used for most of the computational work was purchased with partial financial support of the National Science Foundation.

Chapter 1

INTRODUCTION

Computer-aided research in the field of chemistry has been expanding rapidly over the past few years. The ever-increasing capabilities of today's modern computers are enormous. They are used not only for data storage but also for many computationally intense procedures which were difficult if not impossible to perform before computer assistance (1-4). Chemometrics is just one of the new fields computers have opened up (5,6). This field uses statistics, computers and applied mathematics to solve problems of a chemical nature as illustrated in Figure 1.1.

Quantitative structure-retention relationships (QSRR) are a part of chemometrics which has been the focus of much research (7). The following chapters explain in detail the process of QSRR and discuss the results of two studies in this area.

One of the most common problems in analytical chemistry has been the separation and subsequent analysis of chemical mixtures containing volatile organic compounds. Gas chromatography has proved to be an extremely effective method for doing this. Chromatography is based upon the differential retention of one compound as compared to another as they pass through a column containing a stationary phase. The gas chromatographic stationary phase consists of a material which is polar, non-polar, or perhaps some degree of polarity in between. The solute to be separated and analyzed travels through the column in a solvent called the

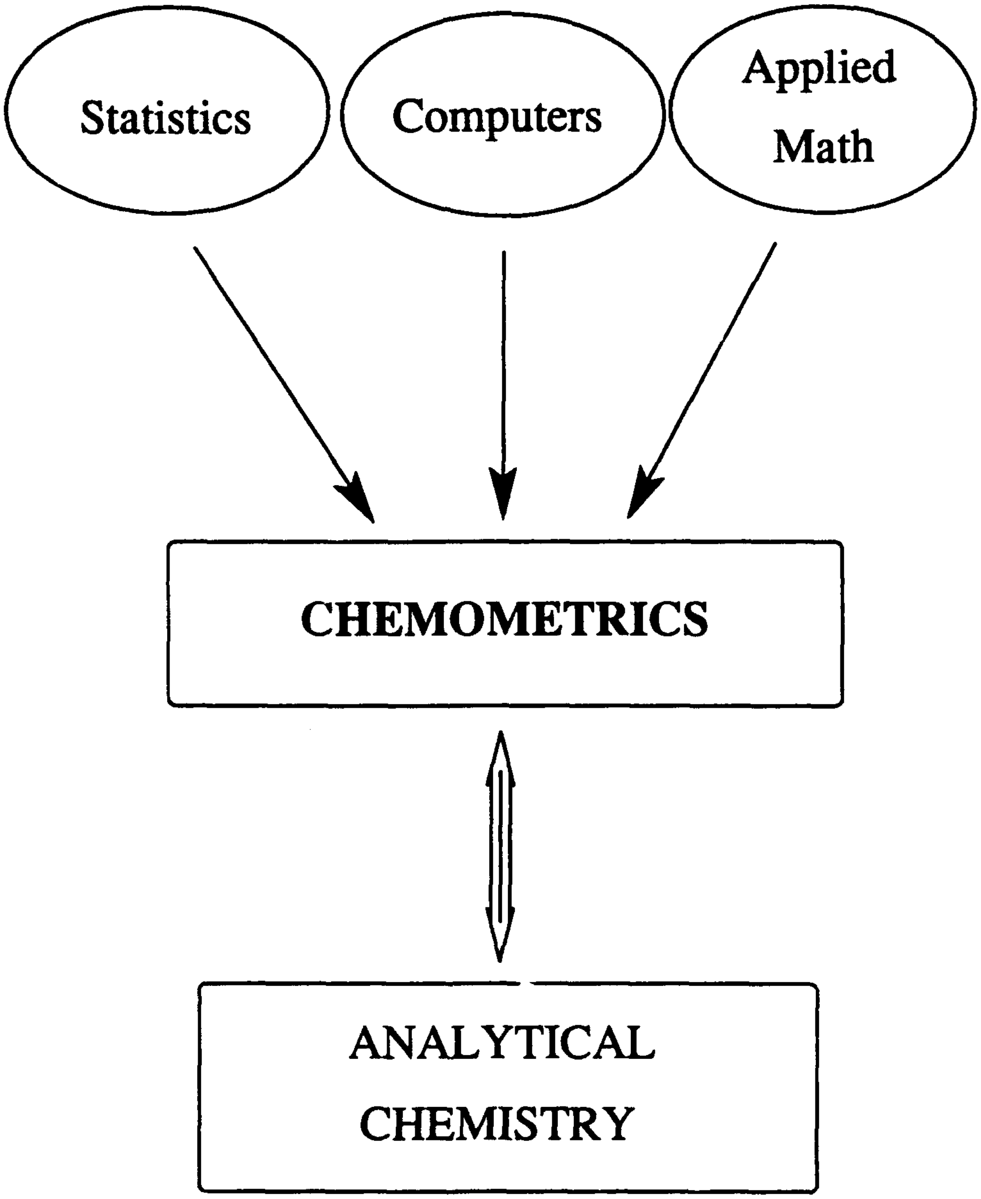


Figure 1.1 The study of chemometrics.

mobile phase. In gas chromatography the mobile phase is called a carrier gas which is usually helium. Th፡ ensuing interactions between the mobile phase, solute and stationary phase are the reasons for the differential retention. While some of the interactions are understood, a detailed understanding of these interactions has not yet been developed.

Some of the most common retention interactions include dipole-dipole, dipole-induced dipole, acid-base and dispersion forces. The nature of the interactions is dependent on the structure of the mobile phase, stationary phase and solute. If the mobile phase and stationary phase are held constant then a relationship between the solute's structure and the column interactions can be developed. Furthermore, a relationship between the solute's retention behavior on a chromatographic column and the solute's structure must also be possible (8,9).

Finding which particular attributes of a compound's structure are important in this relationship is the difficult task; however, once found, this information could be of great use to the analytical chemist. Since differential retention is a means of separating individual compounds from a mixture, a prediction of a solute's retention behavior will be particularly useful when experimentally derived retention data is not available as a standard for comparison. This can easily be the case if the compounds to be studied are extremely toxic, regulated or costly to produce. QSRR data would then be of great value to the chemist seeking to separate and subsequently analyze a mixture.

The first study presented in this thesis models the retention behavior of the 75 polychlorinated dibenzodioxin (PCDD) isomers on five different stationary phases. PCDD isomers are extremely toxic compounds (10). The stationary phases varied in polarity from SP-2100 (non-polar) to SP-2330 (polar). The models obtained for all stationary phases were excellent. Additionally rigorous statistical methods were employed, such as transformations, to obtain only the best models. Outlier analysis was approached from two directions and the methods were compared and contrasted.

The second study involved modeling the retention behavior of the 135 polychlorinated dibenzofuran (PCDF) isomers, which are also toxic (11), on five different stationary phases of varying polarities. Again, excellent results were obtained. Besides the normal statistical analysis, algebraic transformations were used to solve the problems of non-constant variance of the residuals and non-linearity of the regression model. All models maintained their statistical strengths and were internally validated.

Both of these studies were made with homogenous data sets, i.e. similar compounds. The only structural variable was the number of chlorines (one to eight) and the positions of the chlorines on the dioxin or furan backbone. Most homogenous data sets are easily modeled. These data sets are also closed which means there are no more isomers belonging to the PCDDs or PCDFs. Regression models were developed for the entire data set of PCDDs or PCDFs which produced valuable predictions of retention behavior for anyone attempting to separate and analyze these extremely toxic compounds.

References

(1) Vanderginste, BGM. *Trends in Analytical Chemistry*. **1982**, *1*, 210.

(2) McLafferty, F.W.; Stauffers, D.B. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 245.

(3) Campana, J.E.; Jurs, P.C. *Int. J. Mass. Spectr. Ion. Phys*. **1980**, *33*, 119.

(4) Garrison, B.J.; Winograd, N. *Science*. **1982**, *216*, 805.

(5) Sharaf, M.A.; Illman, D.L.; Kowalski, B.R. *Chemometrics*; Wiley Interscience: New York, 1986.

(6) Massart, D.L.; Vanderginste, BGM.; Deming, S.N.; Michotte, Y.; Kaufman, L. *Chemometrics: A Textbook*; Elsevier: Amsterdam, 1988.

(7) Kaliszan, R. *Quantitative Structure-Retention Relationships*; Wiley Interscience: New York, 1987.

(8) Rohrbaugh, R.H.; Jurs, P.C. *Anal. Chem*; **1985**, *57*, 2770.

(9) Hasan, M.N.; Jurs, P.C. *Anal. Chem*. **1988**, *60*, 978.

(10) Karasek, F.; Onuska, F. *Anal. Chem*. **1982**, *54*, 309A.

(11) Hites, R.A. *Acct. Chem. Res*. **1990**, *23*, 194.

Chapter 2

METHODOLOGY

Analyzing multivariate data sets with statistical methods is not unique to chemical problems (1-3). The data is usually broken down into a set of matrices where each member of the data set can be represented by a number or group of numbers which encode information about the member. This is true whether the study involves sociology, meteorology or chemistry. The data set can then be studied by relating the representative group of numbers, or independent variables, to the desired outcome, or dependent variable. A widely used method to obtain a relation is multiple linear regression. In structure-property relationships multiple linear regression (MLR) is used to develop quantitative models relating a set of descriptors, which numerically encode structural information, to a property such as a boiling point (4), or a chromatographic retention index (5).

This chapter details the steps taken to develop models capable of predicting retention data. Quantitative structure-retention relationship (QSRR) studies have been the topic of much work (6,7). The methods used in these studies vary greatly, and the methodology described here is not all inclusive.

QSRR Studies

In any separation scheme involving chromatography there are three important factors which will determine a compound's retention time on a column:

the mobile phase, the stationary phase, and the solute itself. The retention data obtained can be characteristic of a compound thus allowing for the identification of a single compound from a mixture of many compounds. In all QSRR studies presented in this thesis the mobile phase and stationary phase were constant, and therefore it was the difference between the solutes which was analyzed to determine retention. The relationship sought was between the structure of the solutes moving through a chromatographic column and the various retention times of those solutes on the column. A common method in QSRR work involves three basic steps: 1) Collecting of the data set with associated retention data and entry of the structures and retention data into a computer database, 2) numerical encoding of the structural information to generate descriptors which describe the compound's structure, and 3) performing MLR to obtain a model which relates structure to retention while maintaining the model's validity as a predictive device. This method is known as the parametric approach and is shown in Figure 2.1.

The Data Set

Generating a valid data set is crucial in QSRR studies. To start, a set of compounds must have experimentally determined retention data. The retention data is best if it is in the form of retention indexes (8) which can help to eliminate the randomness of the experimental parameters by setting the value of the index relative to a standard compound, usually an alkane. Of course retention indexes are not always available so relative retention times may also be used. A relative retention time (RRT) is the time a compound spends on a chromatographic column relative to an arbitrarily set standard compound. The standard compound's time is usually set

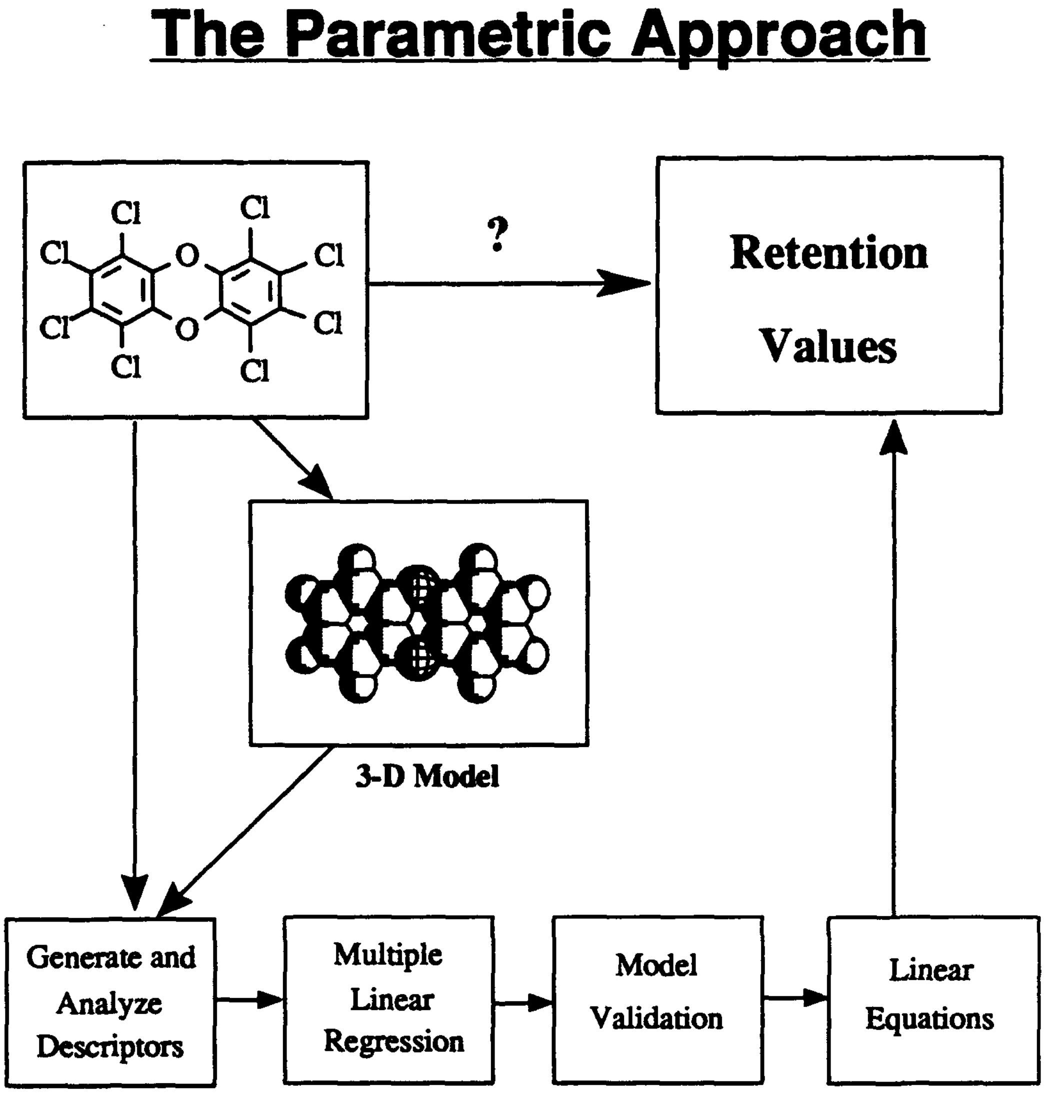


Figure 2.1 The parametric approach for QSRR studies.

to 1.000 minute so any unknown eluting prior to the standard receives an RRT of less than 1.000 minute and any compound eluting after the standard receives a value of more than 1.000 minute.

Data sets can also be relatively diverse in nature. They can consist of compounds containing a wide array of functional groups, various degrees of saturation, and many different atom types. On the other hand, they can be very homogenous such as the PCDDs and PCDFs reported in this thesis. Homogeneous data sets have some advantages. They may be closed, which infers that there is a specific number of compounds belonging to the data set. For example, there are exactly 75 PCDDs and 135 PCDFs. Homogeneous sets contain similar compounds which make structure entry relatively simple. Literature is usually available which contains some experimental retention data; however, not all the data for all isomers of a closed data set are available. Hence the need for QSRR studies.

The experimental data is best if obtained from a single laboratory. This eliminates error between laboratories. Experimental parameters such as column type, column length, temperature program and carrier gas should be reported to make the experiment reproducible. Finally the associated experimental error, if available, is important since the experimental error should never exceed the statistical error in the models. This phenomenon is termed "overfitting" and it interferes with the credibility of the regression models.

Data Set Entry

The data set must be entered into the computer. The structures were entered into a Sun 4/110 workstation via a graphics terminal. The graphics terminal permits the formation of a two-dimensional representation of the compounds. The compounds were stored as a series of bond types (single, double, aromatic, etc.), atom types (C, O, H, Cl) and atom connections in the form of a connection table.

The two-dimensional nature of the connection table limits its ability to describe features such as bond angles, bond lengths, torsional angles and non-bonded interactions. To obtain these features and generate descriptors encoding information about the three dimensional structure, the compounds must be modeled. To accomplish this, the compounds can be modeled classically using methods such as MM2 by Allinger (9,10) or quantum mechanically with a modeling routine such as MOPAC developed by Dewar, et al.(11). Basically both methods create a three-dimensional model of the structure which minimizes the associated strain energy. The strain energy is altered by optimizing various factors such as the bond lengths and bond angles until a potential minimum is located. The three-dimensional model is then viewed graphically to determine if the modeled structure is chemically correct or if perhaps only a local minimum was achieved instead of the desired global minimum. The model can be moved out of a local minimum and the strain re-minimized to the global minimum. It should be noted that no program can do this minimization perfectly and the minimum reached may not be the lowest strain energy possible.

Descriptor Generation

There are three basic classes of descriptors which numerically encode information about the structure of the whole molecule. They are topological, electronic and geometrical descriptors. There is another type of descriptor, but it only encodes information about a specific sector of a molecule. These are called atom-based descriptors and until recently their primary use was for the prediction of ^{13}C NMR chemical shifts (12,13). The following paragraphs characterize the descriptors involved in each class, but descriptors selected for inclusion in specific models will be discussed in greater detail in the appropriate chapter.

Topological Descriptors. These descriptors are calculated from the connection table as discussed earlier. Each structure can be considered as a graph containing a series of nodes and edges. An atom is a node and a bond is an edge. The topological environment is encoded as path lengths, atom types, bond types and others. Molecular connectivity indexes (14) and weighted paths (15) are just two examples. Three-dimensional models are not necessary for the calculation of topological descriptors.

Electronic Descriptors. There are many types of electronic descriptors available. They span the range from Del Re sigma charges (16) through atomic charges by Abraham and Smith (17,18) to extended Hückel (19,20) and CNDO (21) calculations. These descriptors characterize a molecule's autopolarizabilities, partial atomic charges, dipole moment, bonding energies and total energies. Many of these descriptors can be important when describing the polar or non-polar interactions between the solute and the stationary phase. For some descriptors, such as those which rely on throughspace distance interactions, the molecule must be three-dimensionally modeled using the methods discussed earlier.

Geometrical Descriptors. These descriptors numerically describe the three-dimensional nature of a molecule. Again for this class of descriptors a three dimensional model stored as x, y, and z coordinates is essential. Geometrical descriptors include moments of inertia (22), surface area and volume (23-25), and charged partial surface areas (26). The reproducibility of these descriptors is greatly increased if the compounds being studied are fairly rigid and do not possess many low-energy conformations.

Atom-Based Descriptors. Although these descriptors were mainly developed for ^{13}C NMR spectral simulation, they have shown their worthiness in QSRR studies as described in Chapters 3 and 4. Atom-based descriptors are able to numerically encode structural information about a carbon's or a group of carbons' local environment. The carbons of interest may be activated so descriptor generating routines, which calculate topological, electronic and geometrical descriptors, will generate these descriptors for the desired environment only. These descriptors include total average charge (derived from Abraham and Smith) for all heavy atoms one to five bonds away from the atom(s) of interest (18). A heavy atom is any non-hydrogen atom. The variable of one to five bonds away allows the generation of five separate descriptors. Others include heteroatom counts, and Van der Waals throughspace distance interactions (27).

Descriptor Analysis and Model Generation

In QSRR studies over 200 descriptors can be calculated per compound when atom-based descriptors are included. The problem becomes trying to determine a pool of a few information-rich descriptors to submit to regression analysis and model building techniques.

Descriptor Analysis. To find only the best descriptors to use in the study, the descriptors are analyzed both separately and together (28). Separate analysis attempts to delete from consideration all descriptors which have a high number of zero values (usually more than 50%), descriptors which have a high number of identical values (again about 50%), or descriptors with low standard deviations. These descriptors usually contain very little information or cannot be used to resolve the variation in the retention data for the final models. Analyzing descriptors together first eliminates pairwise correlations. As shown in Figure 2.2 columns 3 and 4, although not identical, are exactly correlated (i.e. column 3=2 X column 4). There is redundant information present. Only one of these descriptors needs to be retained. Any pairwise correlations of R>0.90 (R is the multiple correlation coefficient) can usually be eliminated; however, a deleted descriptor can always be switched into the final model to determine if any improvement can be realized.

Secondly, the possibility of some type of multicollinearity also exists. Multicollinearities are determined in three ways. The first is regression of one descriptor against all others, each in turn, to determine the multiple correlation coefficient (MCC or R). An R≥0.95 is usually the cutoff for determining excessive multicollinearity (29).

The second method of multicollinearity detection depends upon the condition index and the related variance decomposition proportions (30). The condition index values can be any number between 1.0 and infinity. The higher the number the greater the possibility of collinearity. Condition indexes are calculated by adding a column of ones to the data matrix where the rows define the compounds and the columns are the descriptors. Next calculate the singular values by taking the square root of the eigenvalues of the correlation matrix. The condition index is then the ratio of the largest singular value to the ith singular value. Belsey, Kuh and Welsch determined a condition index of 30.0 or greater could mean a collinearity problem exists. The variance decomposition proportion is the percentage of total

DATA MATRIX

	DESCRIPTORS				
Compounds	Desc 1	Desc 2	Desc 3	Desc 4	Desc 5
isomer 1	0.012	21.123	2.14	1.07	0.0124
isomer 2	0.054	38.124	8.72	4.36	0.0043
isomer 3	1.287	54.907	5.16	2.58	0.0000
isomer 4	0.564	43.990	2.98	1.49	0.0033
isomer 5	0.500	27.908	4.26	2.13	0.0203
isomer 6	0.400	13.034	1.02	0.51	0.0089
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-

Figure 2.2 Analysis of descriptors.

variance of regression coefficient associated with a particular singular value. For any particular model a condition index greater than 30.0 together with two or more variance decomposition proportions above 0.75 usually infers collinearity problems.

The third method of multicollinearity diagnostics is vector space descriptor analysis (31). Descriptors containing information overlap can be screened out by determining their orthogonality as compared to an initial basis vector. The basis vector is usually the descriptor which is most highly pairwise correlated to the dependent variable. The next descriptor selected will be the one that is most orthogonal to the basis vector and the routine continues until all descriptors are described by a plane angle and distance. Any descriptor which is orthogonal to another will have no information overlap and therefore not correlated. In n-dimensional space n descriptors can be orthogonalized. This process can whittle the final pool of descriptors down to only a few.

Descriptor analysis can be a tedious process. With a pool of over 200 descriptors to start, the process of deleting descriptors can be lengthy. Care must be taken not to submit more than about 25 descriptors at one time for regression analysis, since this would increase the probability of chance correlations (32).

Regression Analysis. The method used for developing models in Chapters 3 and 4 was the linear least squares method (33). In the two studies undertaken regression techniques using interactive regression analysis or forward stepwise regression was chosen (29). Other methods include leaps and bounds (34) and multiple linear regression analysis by progressive deletion (29). These methods are able to generate many different models which all have the capability to fit the retention data to a regression line. To determine the best models, equations are analyzed with respect to the multiple correlation coefficient (*R*), the F-statistic (F), and partial F values for all descriptors. Also analyzed are the standard errors and plots of calculated versus observed values and residual versus calculated values. Another important aspect of model generation was the number of observations to the number of descriptors ratio. If the data set of interest had only 40 experimental observations from which to develop a model for retention behavior, it is obvious that a model with 40 descriptors could perfectly describe the retention behavior. This is not practical nor would the model produced have any predictive power. In order to maintain statistical validity, an observation to descriptor ratio of at least five was maintained. For example, a column with only 15 observations could have a model with no more than three descriptors.

Non-linearity and Non-constant Variance. Generally multiple linear regression analysis can be performed using the generated descriptors without additional scaling or modifying of the dependent or independent variables. Sometimes, however, non-linearity exists between an independent variable and dependent variable. This is easily seen when a plot of the dependent variable versus the independent variable or a plot of calculated values versus observed values is made (Figure 2.3). If it is not readily seen in these plots, which could be the case for a calculated versus observed plot, a residual plot may better illustrate the non-linearity. A residual plot should have a normal distribution around the zero line and display a typical random pattern while a residual plot stemming from a non-linearity problem can look quite different (Figure 2.4). Solving this problem can be accomplished in two ways: (1) transform the independent variable(s) (x-axis) or (2) transform the dependent variable (y-axis). Transformation of the x-axis should always be attempted first as long as the variance of the error terms is generally constant. Transforming the y-axis could bring about the problem of

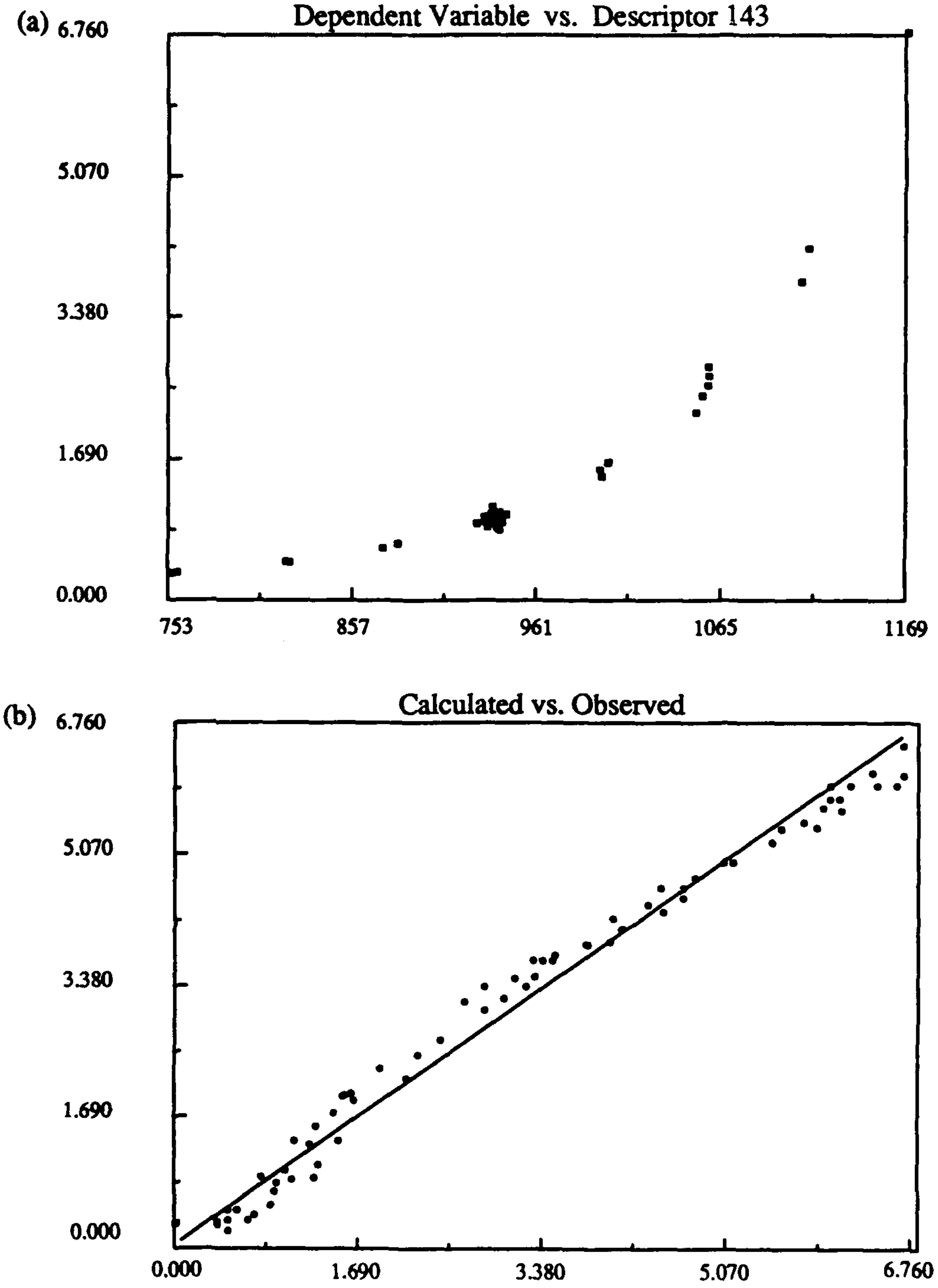


Figure 2.3 Non-linearity of the dependent variable vs. the independent variable (a), and the calculated vs. observed plot (b).

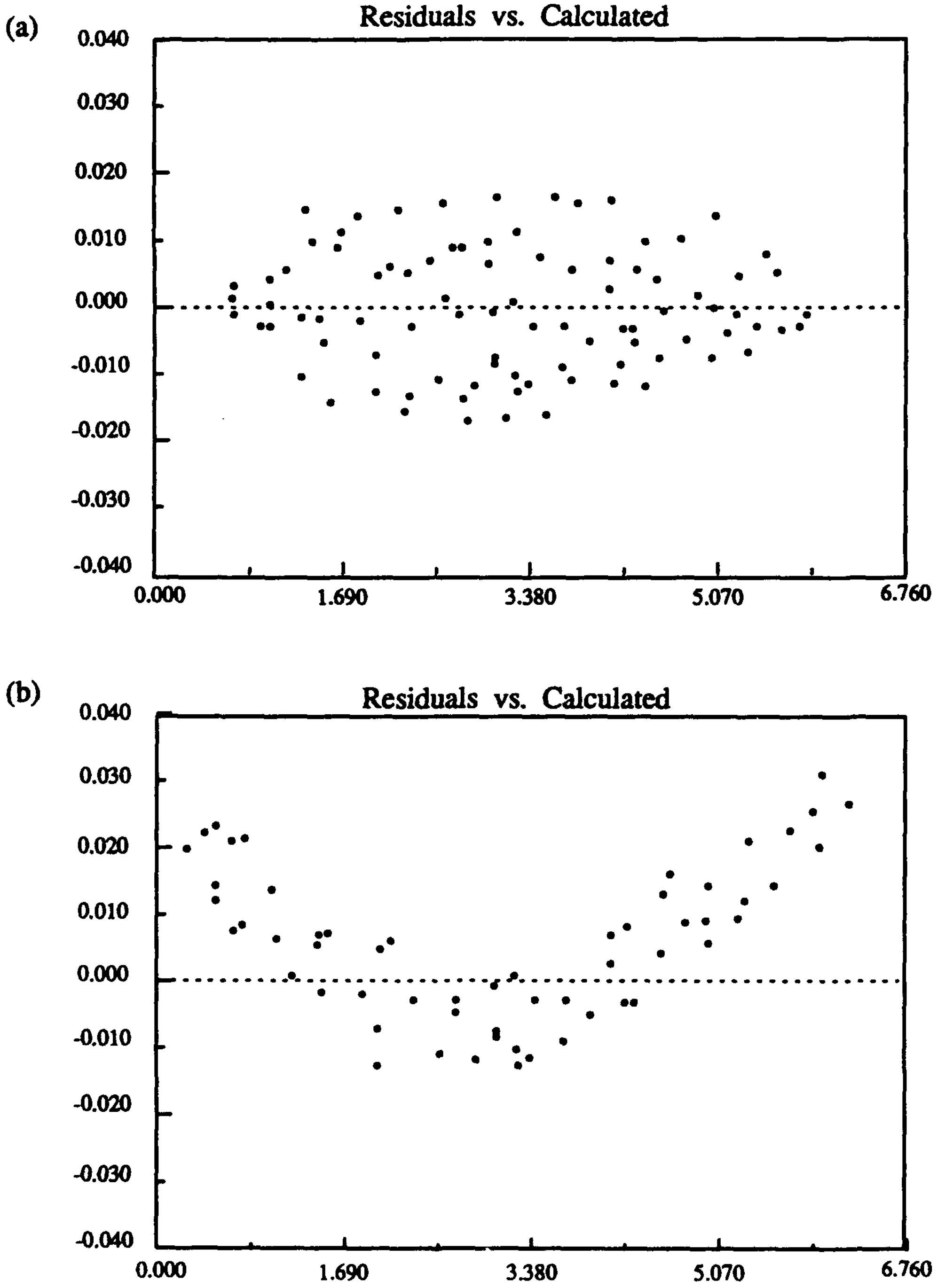


Figure 2.4 Standard residual plots showing normal distribution of error terms (a), and non-linearity of the regression function (b).

non-constant variance of the error terms (29). Transformations to use for typical non-linearities are shown in Figure 2.5. If this does not solve the non-linearity problem then transformation of the y-axis may be attempted; however, residual plots must be examined to determine constancy of variance or the predictive power of the model could be jeopardized. Once the dependent variable is transformed, it must be transformed back to its original form to determine the predicted values for the retention data or the values will be meaningless.

The problem of non-constant error variance is shown in Figure 2.6. Here a transformation of the y-axis is needed, since it is the distribution of the dependent variable which needs to be altered in some way. Different transformations can be tried such as Y_T=log Y, $Y_T=Y^{.5}$, or $Y_T=Y^2$ where Y_T is the transformed dependent variable; however, a general method could be $Y_T=Y^x$ where x=-2, -1.5, -1, -.5, .5, 1.5, 2, etc. This type of general procedure was developed by Box and Cox (35). Sometimes a simple transformation of the dependent variable will solve both a non-linear and non-constant variance problem simultaneously; however, plots of calculated versus observed and residuals must be viewed to confirm the correctness of the transformation. Solving non-linearity and non-constant variance adds to the validity of the final model and usually increases the *R* value while decreasing the standard error of the model.

Outlier Detection

Outliers are observed values which do not fit the generated regression equation well. These points can be graphically detected in most cases in calculated vs. observed plots or residual plots; however, if the outliers fall on the regression

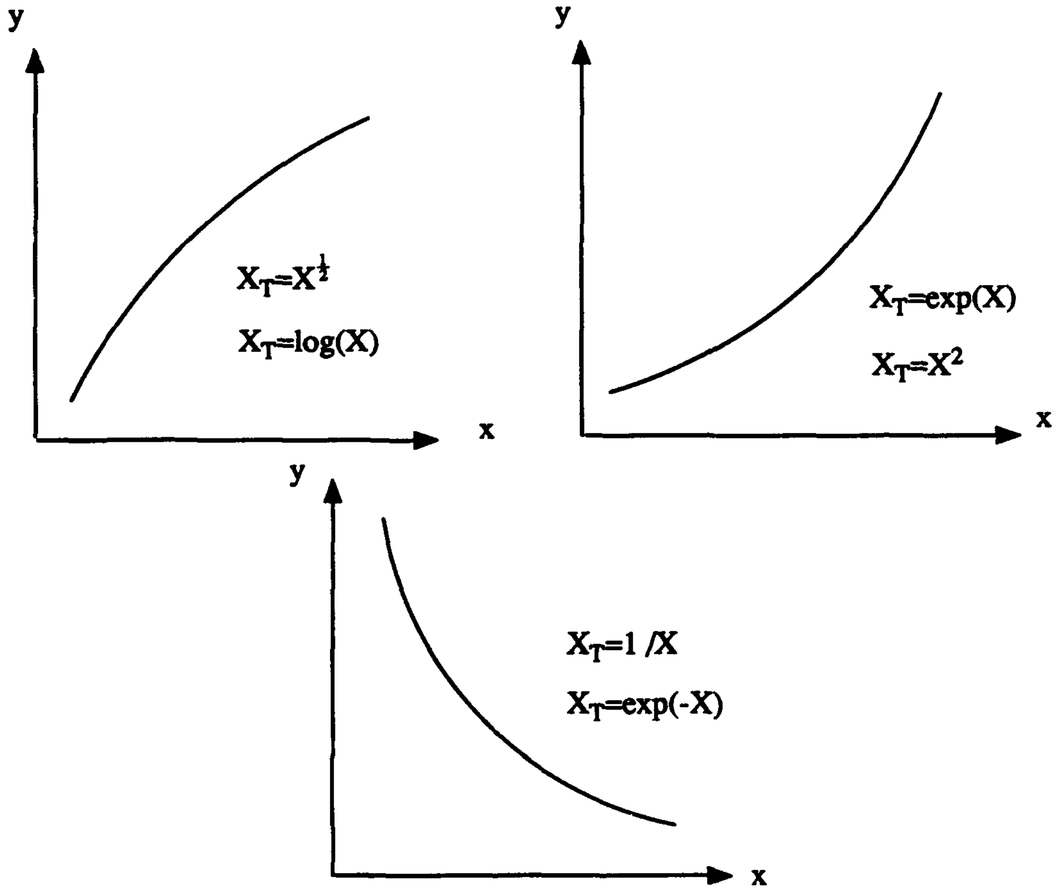
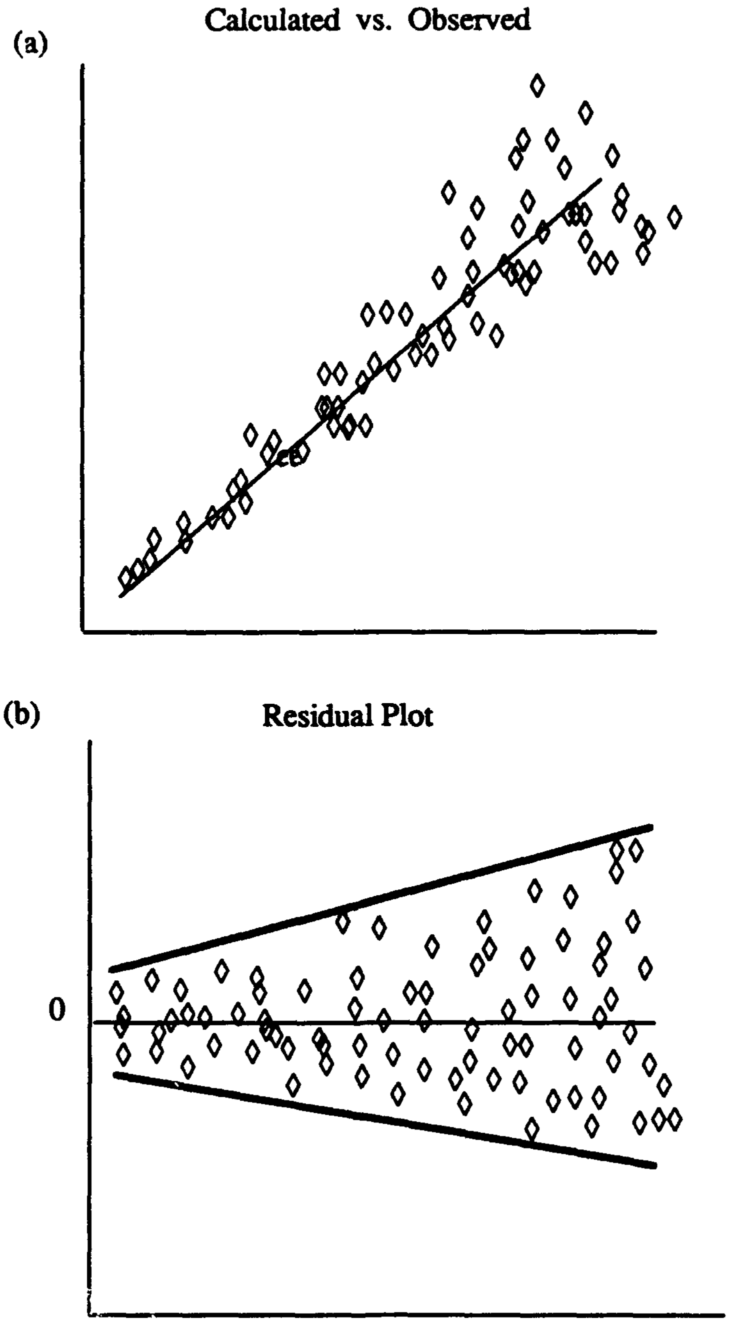


Figure 2.5 Typical transformations of the independent variable or x-axis.


(a) Calculated vs. Observed

(b) Residual Plot

Figure 2.6 Graphical plots showing non-constant variance in (a) the calculated vs. observed plot, and (b) the residual plot.

line at or near the minimum or maximum value of the x-axis, it may be exerting undue pressure on the regression function to conform thereby giving misleading results. This type of outlier may not be detected graphically. Discarding all outliers does not necessarily add to the validity of the model but some form of quantitative and qualitative outlier analysis should be conducted to determine the presence of outliers. Outliers in these studies were determined in two ways: (1) using outlier diagnostics such as leverage values, studentized residuals, DFFITS, Cook's distance, and standardized residuals (29,30), and (2) from Rousseeuw's robust regression analysis (RRA) (36,37).

Diagnostics. The quantitative outlier detection methods listed above are calculated for each of the final models as described in Neter, Wasserman, and Kutner. A limit for each type of test is set, and if the observation in question exceeds the limits of three of these tests, the observation is flagged as an outlier. Although not removed from consideration, it will be subjected to further qualitative and quantitative study to determine possible reasons as to why it was flagged.

Robust Regression Analysis. Robust regression analysis uses a method of least median squares which is not particularly sensitive to the presence of outliers. If this method is compared to the results obtained from interactive regression analysis, the regression coefficients should be nearly identical (or within one standard deviation). If not, there may be outliers present which are affecting the least mean sum of squares regression. RRA will perform the least median squares and display the error in terms of the standard deviation. An observation which lies 2.5 or more standard deviations away from the mean will be flagged as an outlier. In the end a re-weighted least squares is performed on all points not flagged as outliers (outliers are given a weight of zero). Again, any point flagged as an outlier is examined in further detail.

Validating the Model

There are two basic methods of validating a regression model: internal or external.

External Validation. External validation is an excellent method of validation. To accomplish this, some of the original data should be set aside into a prediction set. The remainder of the data is the training set and is used to generate the regression equations. The splitting of the data is done randomly to prevent any statistical bias. Most of the data will be in the training set so as to have the most information available to regress upon. Once the final model has been obtained, predictions are made to see how well the model predicts the retention behavior of the prediction set. In a study with 100 observations, an appropriate training set to prediction set ratio might be 9:1, but can vary according to the amount of data available. One problem with external predictions is that small data sets cannot easily be broken down. For example, a data set with 20 observations can only have four descriptors in the final model in keeping with the minimum 5:1 observations/ descriptor ratio discussed earlier. If some observations are dedicated to a prediction set, not only does the number of possible descriptors in the model decrease but the structural information available from which to generate models also suffers.

Internal Validation. Internal validation in the form of jackknifing is not dependent on the size of the data set and therefore is very suitable for small data sets. In jackknifing, one observation is held out and the model is recalculated. The jackknifed estimate is the predicted value calculated with the new model for the observation being held out. This has also been called the leave-one-out method (38). This calculation is repeated until all observations, in turn, are left out and the jackknifed estimates determined. A jackknifed residual is then the difference between the jackknifed estimate and the experimental observation. Large jackknifed residuals may show possible inconsistencies in the model.

ADAPT System

The ADAPT (Automated Data Analysis and Pattern recognition Toolkit) software contains a series of automated programs which allow the analysis of a wide range of data sets (39). It contains routines for structure entry, molecular modeling, descriptor generation, descriptor analysis (including objective feature selection), regression analysis and validation techniques. It has been proven in QSRR research in the past (5) as well as quantitative structure-activity relationships (QSAR) (40), QSPR (property) (41), and ^{13}C NMR spectral simulation (12,13).

Adapt studies start the same way as outlined in this chapter. The data set is entered into a Sun 4/110 workstation via a graphics terminal using the subroutine UDRAW (42). The initial structure need only be two-dimensional. It is then stored as a connection table. ADAPT can store up to 1000 structures/data set.

The next step, if required, is three-dimensional modeling which ADAPT can also perform with a simple classical method (39) or can be done by more by more rigorous methods (9-11).

The data is then assembled into a worklist containing a training set and a prediction set if possible. Once this is completed, descriptor generation can follow.

All four types of descriptors, topological, electronic, geometrical and atom-based, can be calculated using ADAPT. The descriptors numerically encode information about the structures in the worklist. ADAPT can store 200 descriptors but can calculate well over 200 descriptors. In order to reduce the pool of descriptors to only a few information-rich descriptors, ADAPT routines allow objective feature selection. Pairwise correlations and multicollinearities can be determined as well as vector space descriptor analysis. The remaining descriptors are then submitted to regression analysis and model building.

ADAPT allows regression analysis by progressive deletion, leaps and bounds and interactively. Calculated values and residuals can be plotted with graphical plotting programs. Outlier detection, variance decomposition and jackknifing is performed through ADAPT's data diagnostics generation (DDG) program. After the final models are obtained, predictions of unknown values from the prediction set can be made.

References

(1) Davis, J.C. *Mathematical Geology*. **1970**, *2*, 105.

(2) Fretts, H.C.; Blasing T.J.; Hyaden, B.P.; Kutzbach, J.E. *J. Applied Methodology*. **1971**, *10*, 845.

(3) Treischman, J.S.; Pinches, G.E. *J. Risk Insur*. **1973**, *40*, 327.

(4) Hansen, P.S.; Jurs, P.C. *Anal. Chem*. **1987**, *59*, 2322.

(5) Rohrbough, R.H.; Jurs, P.C. *Anal. Chem*. **1986**, *58*, 1210.

(6) Kalizan, R. *Quantitative Structure-Retention Relationships*. Wiiey Interscience: New York, 1987.

(7) Chan, B.K.; Horvath, Cs. *J. Chromatography*. **1979**, *15*, 171.

(8) Kovats, E. *Chimia*, **1968**, *22*, 459.

(9) Allinger, N.L.; Yul, Y.H.; MM2/MMP2, 85-Force Field (QCPE Program No. 395). Indiana University, IN: Quantum Chemistry Program Exchange, 1985.

(10) Burkert, U; Allinger, N.L. *Molecular Mechanics*; ACS Monograph 177; American Chemical Society: Washington, DC, 1982.

(11) MOPAC, ver 5.0. Quantum Chemistry Program Exchange, QCPE Program No. 445.

(12) Small, G.W.; Jurs, P.C. *Anal. Chem*. **1983**, *55*, 1121-1127.

(13) Sutton, G.P.; Jurs, P.C. *Anal. Chem*. **1989**, *61*, 863-871.

(14) Kier, L.B.; Hall, L.H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.

(15) Randić, M. *J. Chem. Inf. Comput. Sci*. **1984**, *24*, 164-175.

(16) del Re, G. *J. Chem. Soc*. **1958**, 4031-4040.

(17) Abraham, R.J.; Griffiths, L.; Loftus, P. *J. Comput. Chem*. **1982**, *3*, 407-416.

(18) Abraham, R.J.; Smith, P.E. *J. Comput. Chem*. **1988**, *9*, 288-297.

(19) Yates, K. *Hückel Molecular Orbital Theory*; Academic: New York. 1980.

(20) Lowe, J.P. *Quantum Chemistry*; Academic: New York, 1978.

(21) Pople, J.A.; Santry, D.P.; Segal, G.A. *J. Chem. Phys.* **1965**, *43*, S129.

(22) Goldstein, H. *Classical Mechanics*; Addison-Wesley: Reading, MA, 1950, 144-156.

(23) Pearlman, R.S. *In Physical Chemical Properties of Drugs*; Yalkowsky, S.H.; Sinkula, A.A.; Valvani, S.C., Eds; Marcel Dekker: New York, 1980, 321-347.

(24) Pearlman, R.S. *QCPE Bull.* **1981**, *1*, 15-16.

(25) Bondi, A. *J. Phys. Chem.* **1964**, *68*, 441-451.

(26) Stanton, D.T.; Jurs, P.C. *Anal. Chem.* **1990**, *62*, 2323.

(27) Allinger, N.L.; Tribble, M.T.; Miller, M.A.; Wertz, D.H. *J. Am. Chem. Soc.* **1971**, *93*, 1637-1648.

(28) Topliss, J.G.; Edwards, R.P. *J. Med. Chem.* **1979**, *22*, 1238.

(29) Neter, J.; Wasserman, W. Kutner, M.H. *Applied Linear Statistical Models*, 3rd ed; Irwin: Boston, MA, 1990.

(30) Belsey, D.A.; Kuh, E.; Welsch, R.E. *Regression DIagnostics: Identifying Influential Data and Sources of Collinearity*; Wiley Interscience: New York, 1980.

(31) Strang, G. *Linear Algebra and its Applications*, 2nd ed; Academic: New York, 1980.

(32) Stouch, T.R.; Jurs, P.C. *Quant. Struct.-Act. Relat.* **1986**, *5*, 57-61.

(33) Draper, N.R.; Smith, H. *Applied Linear Regression Analysis*, 2nd ed; Wiley Interscience: New York, 1981.

(34) Furnival, G.M.; Wilson, R.W. *Technometrics.* **1974**, *16*, 499.

(35) Box, G.E.P.; Cox, D.R. *J. Royal Stat. Soc.* **1964**, *B26*, 211-243.

(36) Rousseeuw, P.J. *J. Am. Stat. Assoc.* **1984**, *79*, 871-880.

(37) Massart, D.L.; Kaufman, P.J.; Rouseeuw, P.J.; Lerox, A. *Anal. Chem. Acta.* **1986**, *187*, 171-179.

(38) Allen D.M. Technical Report No. 23, 1971; Department of Statistics, University of Kentucky, Lexington, KY.

(39) Stuper, A.J.; Brugger, W.E.; Jurs, P.C. *Computer Assisted Studies of Chemical Structure and Biological Function*; Wiley Interscience: New York, 1979, 83-90.

(40) Lawson, R.G.; Jurs, P.C. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 137-144.

(41) Anker, L.S.; Edwards, P.A.; Jurs, P.C. *Anal. Chem.* **1990**, *62*, 2676-2684.

(42) Rohrbaugh, R.H.; Jurs, P.C. UDRAW (QCPE Program No. 300). Indiana University, IN: Quantum Chemistry Program Exchange, 1988.

Chapter 3

PREDICTION OF GAS CHROMATOGRAPHIC RETENTION DATA FOR POLYCHLORINATED DIBENZODIOXINS

Polychlorinated dibenzodioxins (PCDD) have been the subject of intense study recently (1-3). Their toxicity is well known, and trace analysis of dioxins continues to be performed by many different analytical techniques (4,5). Gas chromatography (GC) is of great use in this area as well as in the separation of dioxin isomers (6). In gas chromatography, Kováts' retention index and relative retention times are normally used for identifying the different isomers. Kováts' retention index is calculated using the equation 1 (7,8) where I_D is the retention

$$I_D = 100N + 100 \frac{\log t_D - \log t_N}{\log t_{N+1} - \log t_N} \qquad (1)$$

index of isomer D, t_D is the corrected retention time of isomer D, t_N is the corrected retention time of the n-alkane standard with a carbon number of N and t_{N+1} is the alkane standard with a carbon number of N+1. Standard N elutes just prior to isomer D.

Relative retention time for the dioxins is the retention time of the isomer in question relative to 2,3,7,8-tetrachloro dibenzodioxin which is the most lethal isomer (1). Some other PCDD studies used the natural abundance ^{13}C 2,3,7,8-tetrachloro isomer as the standard (5,9). This time is reported in minutes.

The PCDDs are a closed data set, meaning there is a limited number of isomers possible. In this case there were only 75 isomers starting with the two mono-substituted isomers and ending with the one octachloro isomer. These isomers can have chlorines attached to carbons 1-4 and/or 6-9 as shown in Figure 3.1. The 75 different isomers are listed in Table 3.1.

The retention of a compound in a chromatographic column depends on the interactions of the solute with the stationary phase. The extent of these interactions is based upon the structural, chemical and electronic properties of the compound. Compounds will usu'lly display unique retention characteristics which will enable separation of each specific compound to its time of retention on the chromatographic column of interest. Retention behavior of PCDDs has been reported for many different stationary phases (10-13).

Experimental Section

To start a QSRR study, a set of relative retention times and/or retention indexes were needed for the dioxins. The data was available from a number of different sources, and although a complete set of all 75 isomers was not available at the time for any particular column type, some data were available for retention on five different stationary phases. Specific column data cross referenced to each compound is also shown in Table 3.1. Column parameters and references are shown in Table 3.2. The exact retention times or retention indexes are also reported in the appropriate reference. The experimental error associated with each column type was not always available. The DB-5 column was taken to have an error of three to seven retention index units from previous retention index work (8). If no error was given it was assumed to be about 1% at the mean of the data points.

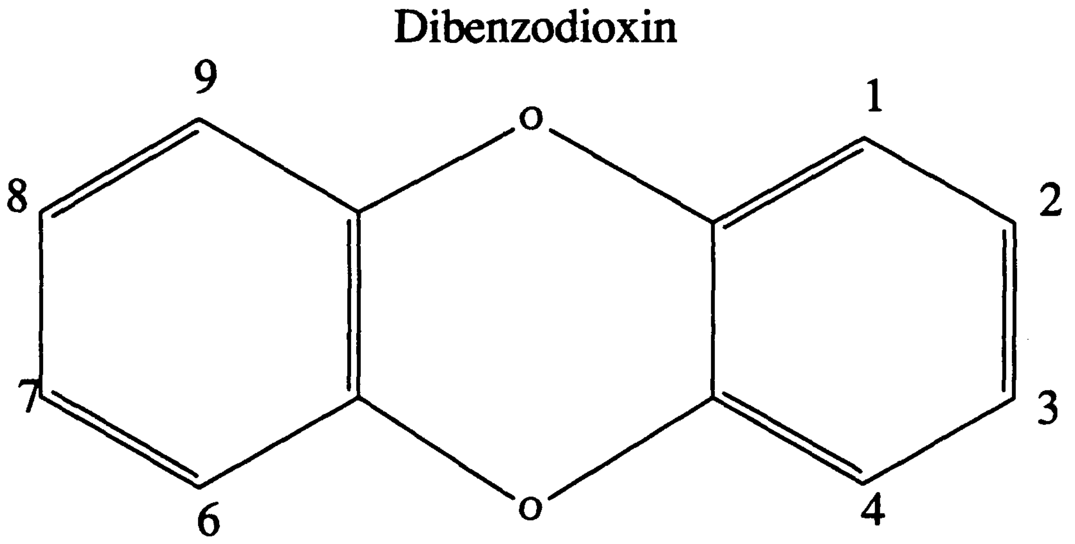


Figure 3.1 Structure of the dibenzodioxin backbone.

Table 3.1 The 75 polychlorinated dibenzodioxin isomers and retention data.

	Column and data type				
	DB-5	SE-54	OV-1701	SP-2330	SP-2100
Isomer	RI	RRT	RRT	RRT	RRT
01 1-Cl					0.293
02 2-Cl					0.299
03 12-diCl					
04 13-diCl					
05 14-diCl					
06 16-diCl					
07 17-diCl					
08 18-diCl					
09 19-diCl					
10 23-diCl	1993				0.433
11 27-diCl	1985				0.424
12 28-diCl	1985				
13 123-trCl					
14 124-trCl	2152				0.600
15 126-trCl					
16 127-trCl					
17 128-trCl					
18 129-trCl					
19 136-trCl					
20 137-trCl					
21 138-trCl					
22 139-trCl					
23 146-trCl					
24 147-trCl					
25 178-trCl					
26 237-trCl					0.651
27 1234-teCl	2379			1.010	0.980
28 1236-teCl	2378			1.020	0.975
29 1237-teCl	2382			1.011	0.985
30 1238-teCl	2382			1.011	0.985
31 1239-teCl	2392			1.068	1.010
32 1246-teCl	2346			1.005	0.910

Table 3.1 (Cont.)

	Column and data type				
	DB-5	SE-54	OV-1701	SP-2330	SP-2100
Isomer	RI	RRT	RRT	RRT	RRT
33 1247-teCl	2340			0.960	0.897
34 1248-teCl	2340			0.960	0.897
35 1249-teCl	2346			1.005	0.910
36 1267-teCl	2408			1.100	1.040
37 1268-teCl	2349			0.977	0.918
38 1269-teCl	2378			1.077	0.972
39 1278-teCl	2400			1.054	1.030
40 1279-teCl	2364			1.021	0.951
41 1289-teCl	2428			1.173	1.090
42 1368-teCl	2290	1.075	1.052	0.876	0.813
43 1369-teCl	2315			0.955	0.852
44 1378-teCl	2340			0.935	0.905
45 1379-teCl	2304	1.082	1.063	0.906	0.833
46 1469-teCl	2341			1.053	0.896
47 1478-teCl	2353			0.994	0.928
48 2378-teCl	2386	1.125	1.106	1.000	1.000
49 12346-peCl					
50 12347-peCl	2573				1.540
51 12367-peCl	2604				
52 12368-peCl		1.215	1.189		
53 12369-peCl					
54 12378-peCl	2587	1.253	1.229		1.630
55 12379-peCl		1.225	1.203		
56 12389-peCl	2623				
57 12467-peCl					
58 12468-peCl	2501	1.192	1.170		
59 12469-peCl					
60 12478-peCl		1.220	1.196		1.460
61 12479-peCl	2501	1.192	1.170		
62 12489-peCl					
63 123467-heCl	2812				
64 123468-heCl	2742				
65 123469-heCl					

Table 3.1 (Cont.)

			Column and data type			
		DB-5	SE-54	OV-1701	SP-2330	SP-2100
	Isomer	RI	RRT	RRT	RRT	RRT
66	123478-heCl	2781	1.411	1.370		2.540
67	123678-heCl	2788	1.409	1.363		2.650
68	123679-heCl		1.337	1.338		2.420
69	123689-heCl		1.337	1.338		
70	123789-heCl		1.432	1.395		2.760
71	124679-heCl	2713				2.220
72	124689-heCl	2713				
73	1234678-hpCl	2994	1.659	1.588		4.180
74	1234679-hpCl	2949				3.780
75	12346789-ocCl	3196				6.760

RI--Retention Index
RRT--Relative Retention Time

Table 3.2 Column parameters.

Manufacturer[a]	J&W	Supelco	H-P	Supelco	H-P
Column Type[b]	DB-5	SE-54	OV-1701	SP-2330	SP-2100
Column I.D.(mm)	0 .25	0.3	0.3	0.25	0.2
Column Length (m)	60	25	20	60	50
Film Thickness (µm)	0.25	0.10	0.10	0.2	NA
Carrier Gas	He	He	He	H_2	He
Data[c]	RI	RRT	RRT	RRT	RRT
Temperature	170 (1 min)	60-260	60-260	100-180 @2°/min	225
Program (°C)	170-340 @2°/min	10°/min	10°/min	180-260 @5°/min	Const.
Reference	(10)	(11)	(11)	(12)	(13)

a) Columns obtained from J&W Scientific, Supelco, or Hewlett Packard.

b) DB-5 is a methyl silicone column with 5% phenyl substitution.
SE-54 is similar to DB-5. It is being replaced by DB-5.
OV-1701 is 86% dimethyl polysiloxane and 14%cyanopropyl phenyl.
SP-2330 is 90% bis cyanopropyl and 10%cyanopropyl phenyl polysiloxane.
SP-2100 is 100% dimethyl polysiloxane.

c) RI=Retention Index calculated using equation 1.
RRT=Relative Retention Time which is relative to 2,3,7,8 tetrachloro dibenzodioxin which = 1.000 min. except for SE-54 and OV-1701 which are relative to 5-chloro-2-(2,4-dichlorophenoxy)-anisole which =1.000 min. Absolute retention times for the standard on SE-54 and OV-1701 is 18.14 min. and 17.54 min. respectively.

The QSRR study consisted of four stages as described in Chapter 2: 1) Entry, modeling and storage of the structures as a series of x, y, and z coordinates and the corresponding retention data; 2) generation of descriptors; 3) multiple linear regression analysis; and 4) model validation techniques.

Entry, Storage and Molecular Modeling

The PCDD structures were entered into the ADAPT system using the UDRAW subroutine as described in Chapter 2. The structures were first stored as a connection table of atom types, bond types and bond connections. Three dimensional models were generated using the classical molecular modeling routine of ADAPT. Although this modeling program is not as robust at minimizing strain energies as Allinger's MM2 (14,15), or MOPAC (16), it was determined that the dioxins were all fairly planar and this was confirmed by modeling a few selected isomers with MM2 and AMPAC. All three routines showed the dioxins as planar so the remaining isomers were modeled with the basic ADAPT program.

Descriptor Generation

The structure of a molecule can be described by a set of numerical values. The values can then directly represent the properties of the molecule. The descriptors calculated in this study were topological, electronic and geometrical, but in contrast to other studies of QSRR (17,18), this study saw the introduction of atom-based descriptors to describe the environment surrounding the bridgehead carbons as shown in Figure 3.2. These descriptors proved to be of great value.

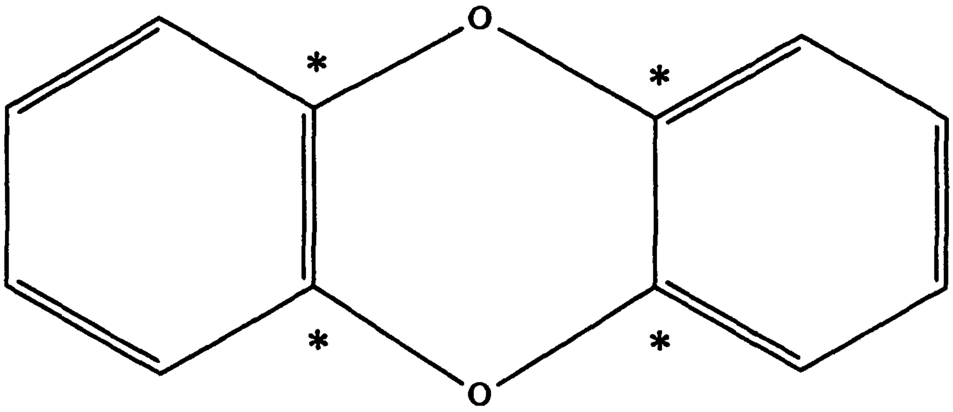


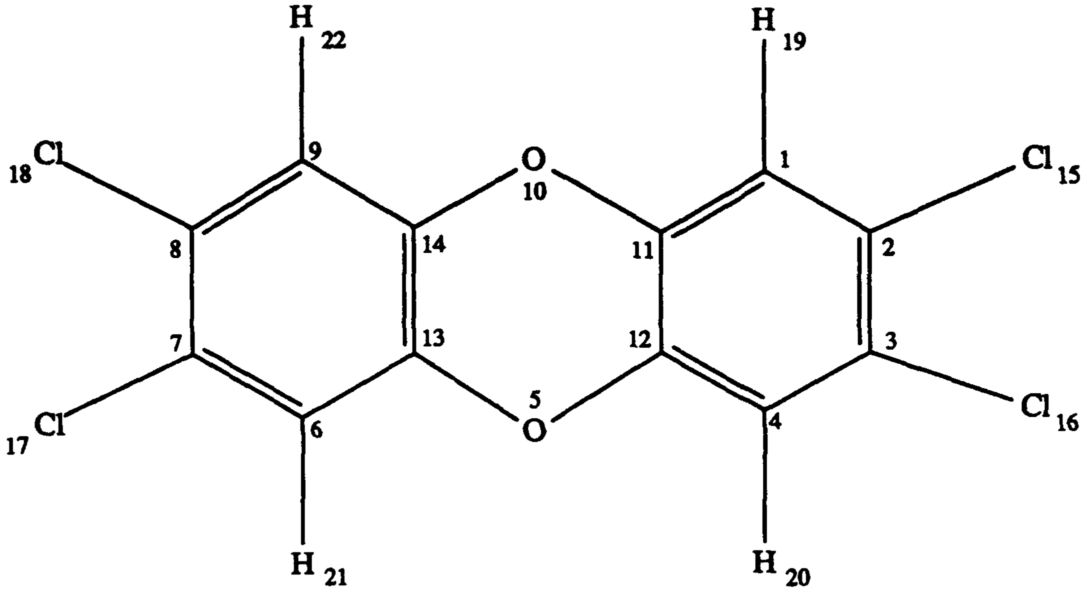
Figure 3.2 Atom-based descriptors describe the environment surrounding the bridgehead carbons (*).

Topological Descriptors. The topological descriptors used were path, cluster and general shape indices. All of these descriptors can be generated from a two-dimensional model of the compound which is stored as a connection table. The $^7\chi_{CH}$ descriptor is a chain path descriptor (19). The shape descriptors are Kappa indices as described by Kier (20). These descriptors have been used before in other studies and correlate well with structure-retention relationships.

Electronic Descriptors. Three electronic descriptors were important for predicting the retention behavior of the PCDDs. The first was a sum total of all partial negative charges of the molecule as calculated with an equation from Abraham and Smith (21,22). This was accomplished using an ADAPT program developed by Dixon (23). The other two descriptors involved simple Hückel theory (24). The simple Hückel calculations led to the total energy of the molecule and the electron density minimum. These two values are related to the number of chlorines present and the positions they occupy. Compounds with chlorines positioned near the oxygen have a greater retention index or longer relative retention time (see data in Table 3.1), a greater total energy and higher strain energies.

Geometrical Descriptors. These descriptors must be calculated using the three-dimensional x, y and z coordinates of the compounds. This was why the time was taken to model (minimize the strain energy of) the dioxins carefully. Many of these descriptors utilize the throughspace distance or bond length between atoms in their calculations. Others use the surface area or volume of a molecule to calculate a charged partial surface area (25). These types of descriptors have been valuable in other studies (25,26). Two whole molecule geometrical descriptors were calculated for this study. The length-to-breadth ratio (L/B) is the minimum ratio of the molecule's length compared to its breadth (27). It is effectively encoding the positions of the chlorines. For example if a tetrachloro isomer has chlorines at the 1, 4, 6, 9 positions (see Figure 3.1) its L/B ratio would be substantially smaller than the 2, 3, 7, 8 isomer. The other geometrical descriptor was based upon the symmetry of the molecule. The descriptor determines the number of unique atoms in a molecule and divides this by the total number of atoms to form a symmetry index. The descriptor uses throughspace and bond distances to determine an atom's uniqueness.

Atom-Based Descriptors. At first glance, the position of the chlorine atoms seemed to influence the magnitude of the retention index or retention time. It became necessary to investigate this fact and look for a descriptor which could encode the necessary chlorine position information. Furthermore, when the chlorines were near the oxygen the retention values were greater on all column types, polar or non-polar. Therefore descriptors which could actually describe the environment near the oxygen topologically, electronically and geometrically would be of great use. Atom-based descriptors similar to those used in ^{13}C NMR spectral simulation work were used. Three atom-based descriptors were used in this study. The environment of the bridgehead carbons was determined to be the area of concern for descriptor calculations. The bridgehead carbons are the nearest neighbors to the oxygen and by describing the environment of the bridgehead carbons, the environment of the oxygens would also be partially described. The first descriptor was based upon Extended Hückel Theory (24,28). The average Hückel charge of all heavy atoms (non-hydrogen) three bonds away from the bridgehead carbons was calculated. The second descriptor was derived from Dixon's charge program based on the partial atomic charges for each atom as shown in Figure 3.3. The descriptor is a step atomic charge descriptor calculated using methods described by Abraham and Smith (21,22). The last atom-based descriptor calculated was the



ATOM	TYPE	CHARGE
1	C	-.0492
2	C	+ .0738
3	C	+ .0738
4	C	-.0492
5	O	-.1751
6	C	-.0492
7	C	+.0738
8	C	+.0738
9	C	-.0492
10	O	-.1751
11	C	+.0804
12	C	+.0804
13	C	+.0804
14	C	+.0804
15	Cl	-.1022
16	Cl	-.1022
17	Cl	-.1022
18	Cl	-.1022
19	H	+.0848
20	H	+.0848
21	H	+.0848
22	H	+.0848

Figure 3.3 Partial atomic charges for each atom of 2,3,7,8 tetrachloro dibenzodioxin.

Van der Waals energy of the bridgehead carbons interacting with all heavy atoms one bond away from the bridgehead atoms. For each atom-based descriptor, the value for each of the four bridgehead carbons was calculated and then the average of the four was taken as the descriptor value. These descriptors encode information which describes the bridgehead environment as a whole. These atom-based descriptors also contain topological, electronic and geometrical information, but where whole molecule descriptors encode structural information about the total molecule, these descriptors look at a specific area or even a specific atom's environment.

Regression Analysis

More than 200 descriptors were calculated for each of the 75 PCDD isomers; however, not all of these descriptors can be used in the model. To delete descriptors from the list, about 25 descriptors at a time were subjected to analysis in which descriptors encoding nearly the same information (i.e. highly correlated, R>0.90) were deleted. This was described in Chapter 2 as objective feature selection. High pairwise correlations can influence the validity and predictability of the final model. Objective feature selection was performed and a pool of about 40 descriptors remained. These remaining descriptors were entered into a vector space descriptor analysis program to determine which descriptors had the highest information content while also displaying any remaining multicollinearities. The dependent variables were the observed retention data for each of the stationary phases. The top few descriptors containing the most information were then regressed upon each other to determine if any multicollinearities still existed. The variance decomposition proportions and related condition indexes were also examined as discussed in Chapter 2. As a result of these tests, a total of 11 descriptors could be used to predict the retention of the PCDDs on five different stationary phases. More than 190 descriptors were deleted in these steps. The final 11 descriptors (Figure 3.4) were then submitted to interactive regression analysis to determine the best possible model for each column.

Results and Discussion

Models were developed for each column that contained anywhere from two to five descriptors depending upon the number of observations available to use in the modeling process. Since not more than 41 observations were available out of a maximum of 75 for any one column, no observations were held out to use as a prediction set as described in Chapter 2. Therefore, the entire set of observations was used to develop models and the models were internally validated.

The DB-5 stationary phase had the greatest number of experimental observations available (N=41). This was also the only data set where the author reported a retention index instead of a relative retention time. The best equation developed by interactive regression analysis for the DB-5 is given in equation 2. The coefficients for the model are listed in the order of information content with respect to the dependent variable. For instance, TOAC 3 was the most highly correlated with the retention index so it was selected first. The numbers with the coefficients are the 95% confidence intervals for each of the coefficients. The standard error of 9.84 retention index units corresponds to an error of approximately 0.8% of the range. The experimental error was not available in (10), but from other

Descriptors Submitted to Regression Analysis

AVHC 3[a] Average Hückel charge for all heavy atoms 3 bonds away from the bridgehead carbons. The value is the average of all four carbons.

TOAC 3[a] Sum of the absolute values of the atomic charges for all heavy atoms 3 bonds away averaged over the four bridgehead carbons.

CXVD 1[a] Van der Waals energy of the bridgehead carbons interacting with other heavy atoms. Only 1-4 interactions or greater are included.

$^{7}\chi_{CH}$[b] Simple 7th order chain ring path.

KAPA 3[b] Shape index relating atom types.

KAPA 6[b] Shape index relating atom and bond types.

QNEG[c] The charge on the most negative atom.

EDMN[c] Electron density minimum. This is the sum of the contributions of the atomic orbitals to each of the molecular orbitals multiplied by the number of electrons in the molecular orbitals.

ETOT[c] Total Hückel energy of the system which is a sum of the energies of all occupied molecular orbitals.

L/B[d] Minimum length-to-breadth ratio.

SYMM 35[d] Encodes symmetry by calculating an index defined as the number of unique structural atoms divided by the total number of atoms.

(a) Atom-based.

(b) Topological.

(c) Electronic.

(d) Geometrical.

Figure 3.4 The final pool of descriptors submitted to regression analysis.

$$\begin{aligned} RI = {} & 980.1 \pm 8.952 \text{ (TOAC 3)} + \\ & -426.7 \pm 28.09 \text{ (KAPA 3)} + \\ & 1278 \pm 140.7 \; ({}^{7}\chi_{CH}) + \\ & 102.8 \pm 11.32 \text{ (L/B)} + \\ & 1087 \text{ (INTERCEPT)} \end{aligned} \qquad (2)$$

$N=41 \qquad s=9.84\ (0.4\%) \qquad R=0.999 \qquad F=6938$

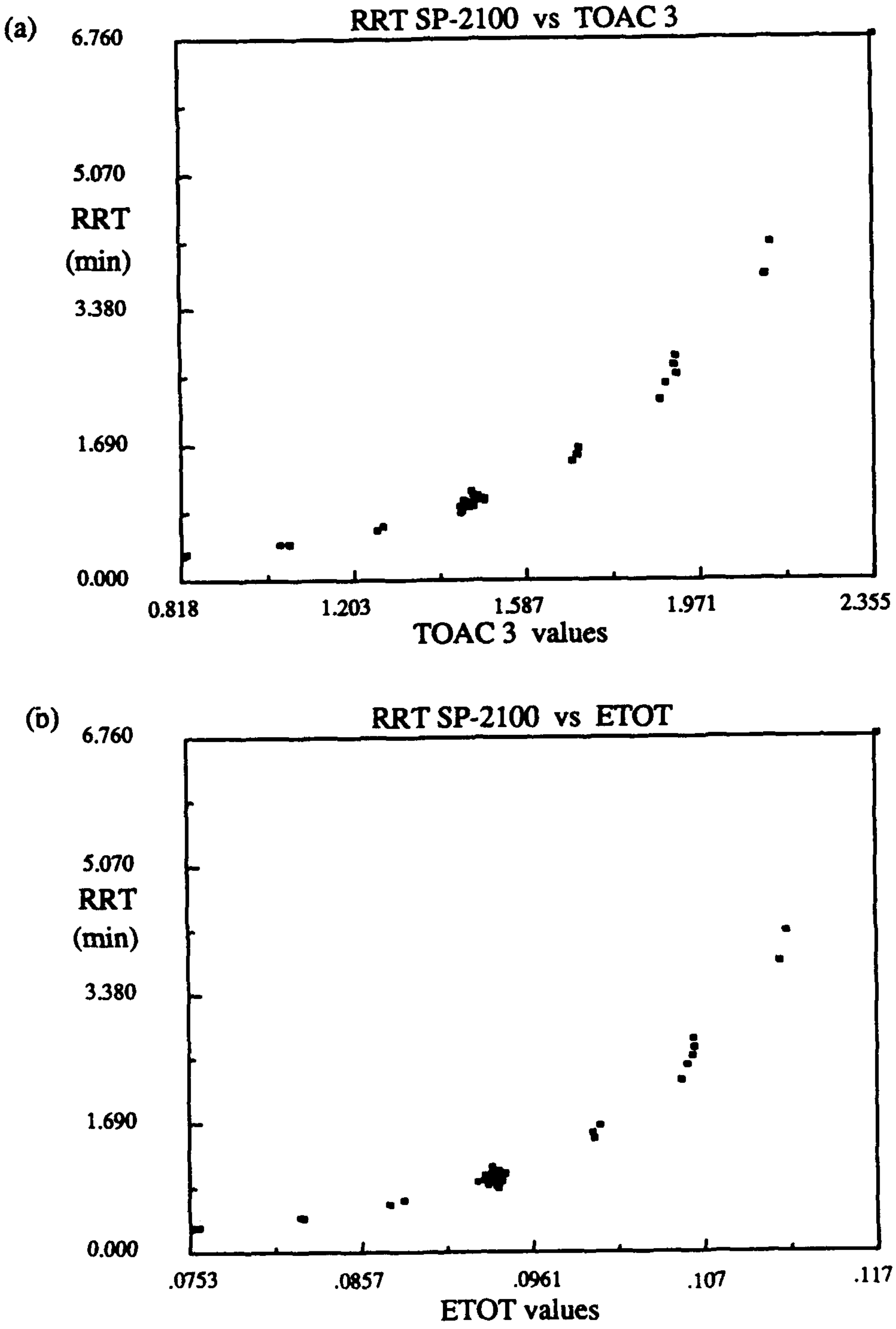
work done with retention index equations the reproducibility associated in this type work is usually three to seven index units (8). A recent study modeling dibenzofurans on DB-5 also found experimental errors in this range (29). The model presented here does not overfit the data but it does have the ability to predict fairly close to the experimental error. The low standard error and high F value coupled with a P value less than 0.0001 demonstrates the superb fit of the calculated values.

The same procedures were used to model the other four stationary phases. The number of descriptors used to model any one of the columns never violated the rule of one descriptor for every five observations. On the columns with only 15 observations only three descriptors could be used. This makes it difficult to obtain good models. The results of modeling are shown in Table 3.3. Only descriptors which were significant for a particular column are shown with coefficients.

One of the stationary phases, SP-2100, was extremely hard to model at first. Upon graphing the dependent variable versus the independent variable some interesting relationships became apparent. Figure 3.5 shows the plots of relative retention times versus the values for TOAC 3 and the values for ETOT. The non-linear relationship was clearly evident and different transformations on the

Table 3.3 Summary of models.

			Column type		
Descriptor	DB-5	SE-54	OV-1701	SP-2330	SP-2100
AVHC					
$^{7}\chi_{CH}$	1278±141			1.93±0.45	-27.08±3.70
KAPA 3	-426.7±28.1			-0.772±0.068	
KAPA 6		2.652±1.116			
QNEG				-16.63±2.29	
TOAC 3	980.1±8.95	0.932±0.088			
EDMN				3.23±0.77	37.76±10.66
ETOT			2.586±0.110		12.68±1.94
L/B	102.8±11.3		0.184±0.034		2.06±0.041
CXVD			-0.335±0.077		-4.00±1.26
CONST.	1087	-5.81	-45.16	-4.44	-248.7
R	0.999	0.980	0.996	0.981	0.966
s	9.84(0.4%)	0.034(2.5%)	0.014(1.1%)	0.014(1.3%)	0.346(9.8%)
F	6938	142	473	111	91
N	41	15	15	22	39



RRT SP-2100 vs TOAC 3

RRT SP-2100 vs ETOT

Figure 3.5 Plots of relative retention times for SP-2100 versus (a) the values for TOAC 3 and (b) the values for ETOT.

dependent variable (x-axis) were attempted as discussed in Chapter 2. At first a simple square (X^2) function and antilog (10^x) were tried, but nearly the same non-linear relationship persisted. This was due to the narrow range of values of the independent variable which simple squaring would not correct. Upon further investigation it was determined that a transformation of the dependent variable or y-axis could eliminate the non-linearity; however, three precautions must be adhered to. First, the transformed y-axis must be plotted against the available descriptors and the plots should not show any non-linearities. This was performed and no non-linearities were found. Second the new y-axis must be modeled or regressed again with the descriptor pool to obtain a new model with different regression coefficients. This was done and the results are shown in Table 3.4. Third, after transforming the y-axis the error variance associated with the model must be analyzed to determine if it is constant. This was accomplished graphically as explained in Chapter 2. The residual plot in Figure 3.6 after transformation of the y-axis showed no evidence of non-constant variance. The new model was plotted as a calculated versus observed plot and no curvilinear relationship was found (Figure 3.7). The actual transformation is shown in equation 3 where Y_T is the new dependent variable. A factor of 100 was used as a scaling factor. An R=0.999 showed quite an improvement. The most important step was to be sure the transformation did not interfere with the constancy of the error variance.

The model calculates the log of the retention time instead of the actual retention time so any predictions must be transformed back to retention times by taking the antilog of the raw predictions and dividing by the scaling factor. The reason the non-linearity existed for only the SP-2100 column and not the others can be found upon examination of the column data in Table 3.2. The experimental

Table 3.4 Coefficients for the SP-2100 model after transforming the y-axis.

$$\begin{array}{ll} \text{RRT=} & 3.813 \pm .064\ \text{(ETOT)} \\ & -0.442 \pm .036\ \text{(KAPA 3)} \\ & \ \ 0.144 \pm .018\ \text{(L/B)} \\ & -0.640 \pm .170\ \text{(AVHC 3)} \\ & -65.60\ \ \text{(Intercept)} \end{array}$$

R=0 .999

s = 0.015(0.7%)

F= 3766

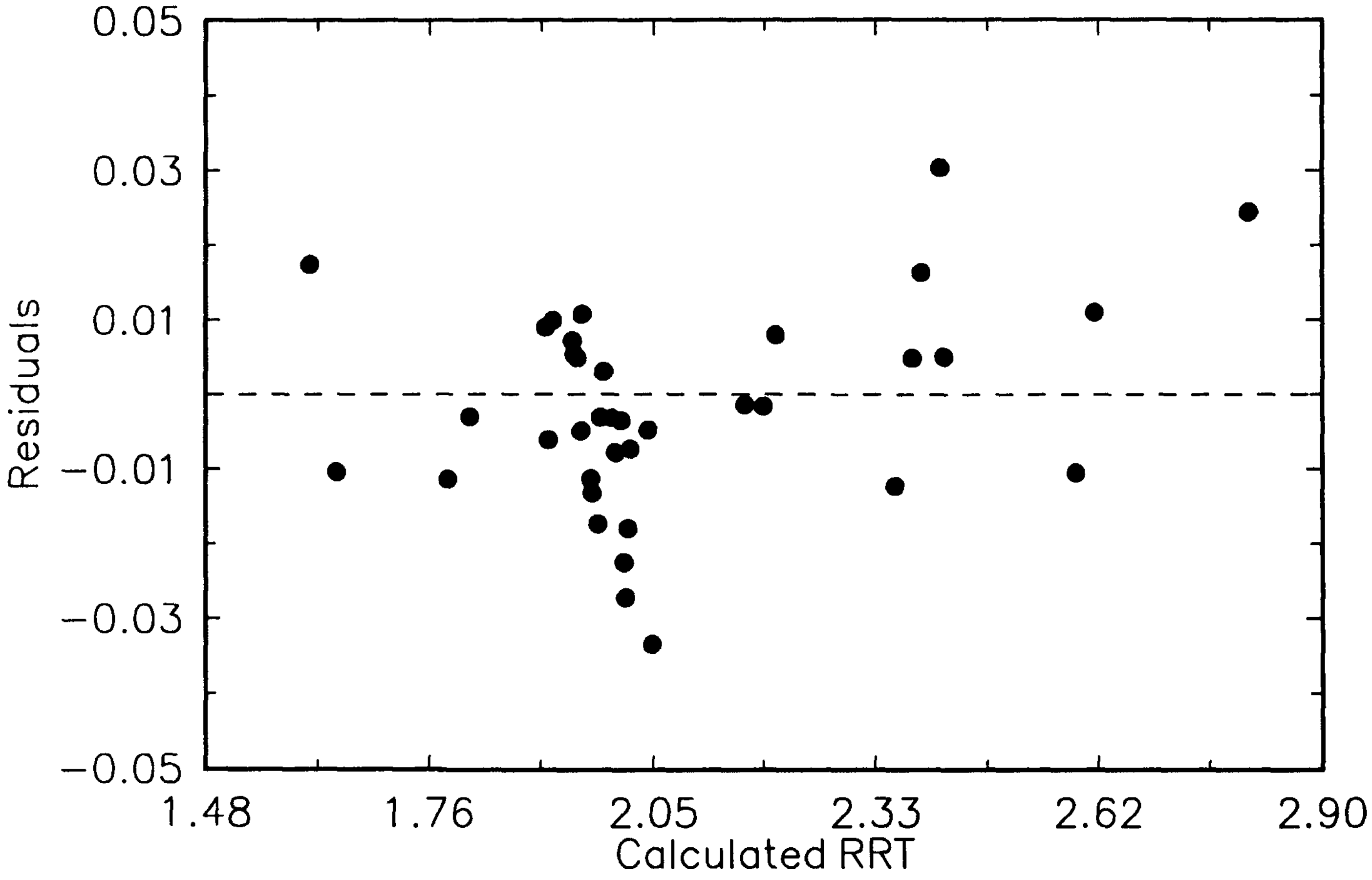


Figure 3.6 Residual plot of SP-2100 column after transformation.

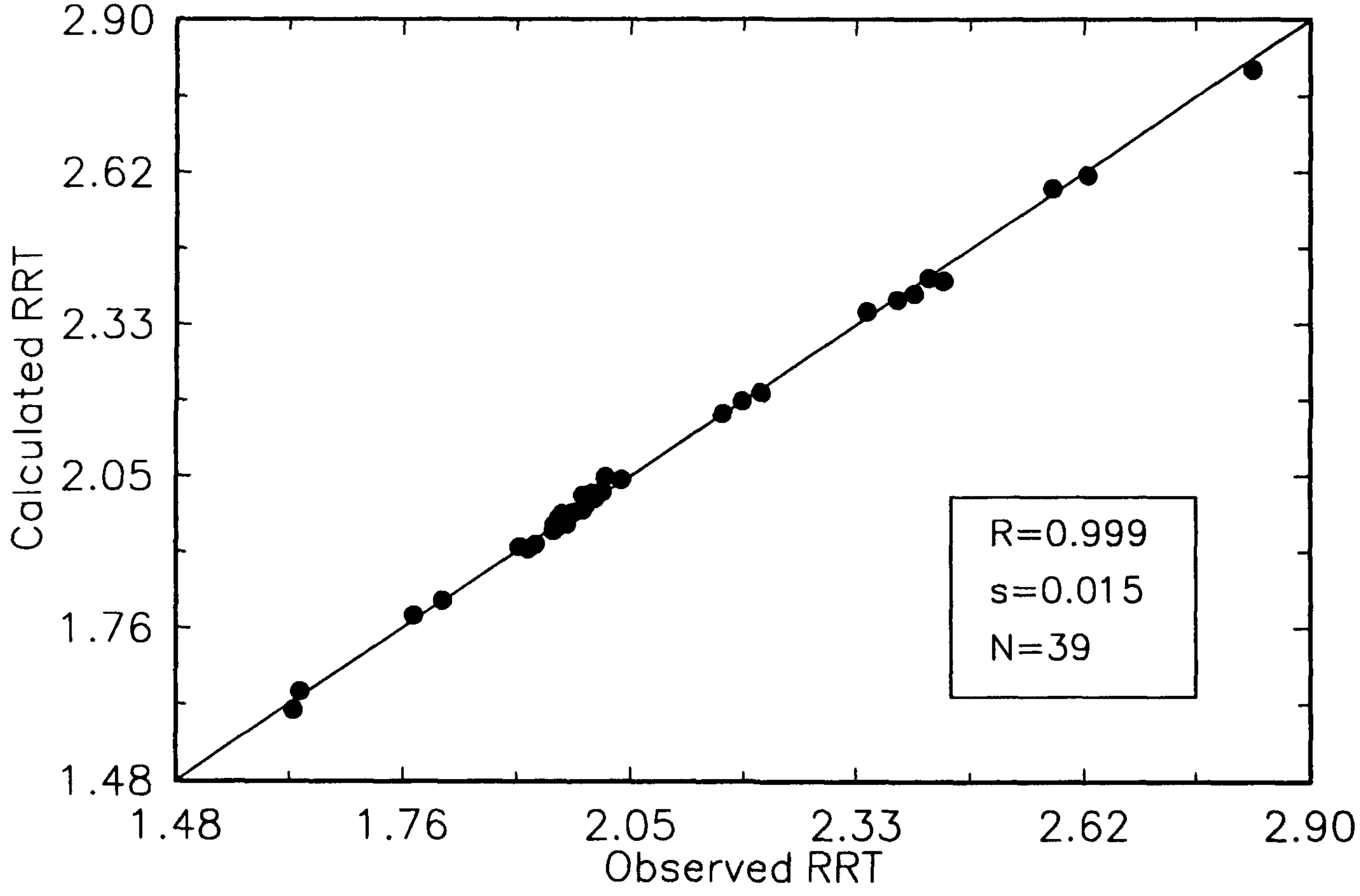


Figure 3.7 Calculated vs. observed plot for SP-2100 column after transformation.

$$Y_T=\log(100\,(RRT)) \qquad (3)$$

parameters show that SP-2100 was the only column where the temperature was held constant and no programming was attempted. When the dependent variables of the other columns were plotted against the descriptor values no non-linearity was observed. This non-linear behavior for isothermal GC data has been documented before (10).

Outlier Detection

Outliers tend to be poorly fitted points which for some reason cannot be brought back into the fitted region without compromising the validity of the model.

Two different methods were used to check for outliers as discussed in Chapter 2. The data diagnostics generation (DDG) routine of ADAPT and robust regression analysis (RRA) (30,31) were compared and contrasted against each other to determine which compound for each column could be considered as outliers. DDG uses a number of tests (see Chapter 2) such as DFFITS, leverage values, Cook's distance and studentized residuals to determine outliers. Generally, if the cutoff values for three of the tests were exceeded the point was taken to be an outlier. The outliers determined by DDG for each column are shown in Table 3.5. RRA uses a least median squares approach instead of a least mean squares method. Any point which has a residual greater than 2.5 times the standard error was determined to be an outlier. Outliers determined by RRA are shown in Table 3.6.

Table 3.5 Outliers determined by DDG.

Isomer	DB-5	SE-54	OV-1701	SP-2330	SP-2100
1-monoCl					X
2-monoCl					X
1234-teCl				X	
1267-teCl					
1269-teCl					
1278-teCl					
1289-teCl				X	
1469-teCl				X	
1478-teCl					
2378-teCl			X		X
12378-peCl					
1234678-hpCl		X	X		
12346789-ocCl	X				

Table 3.6 Outliers determined by RRA.

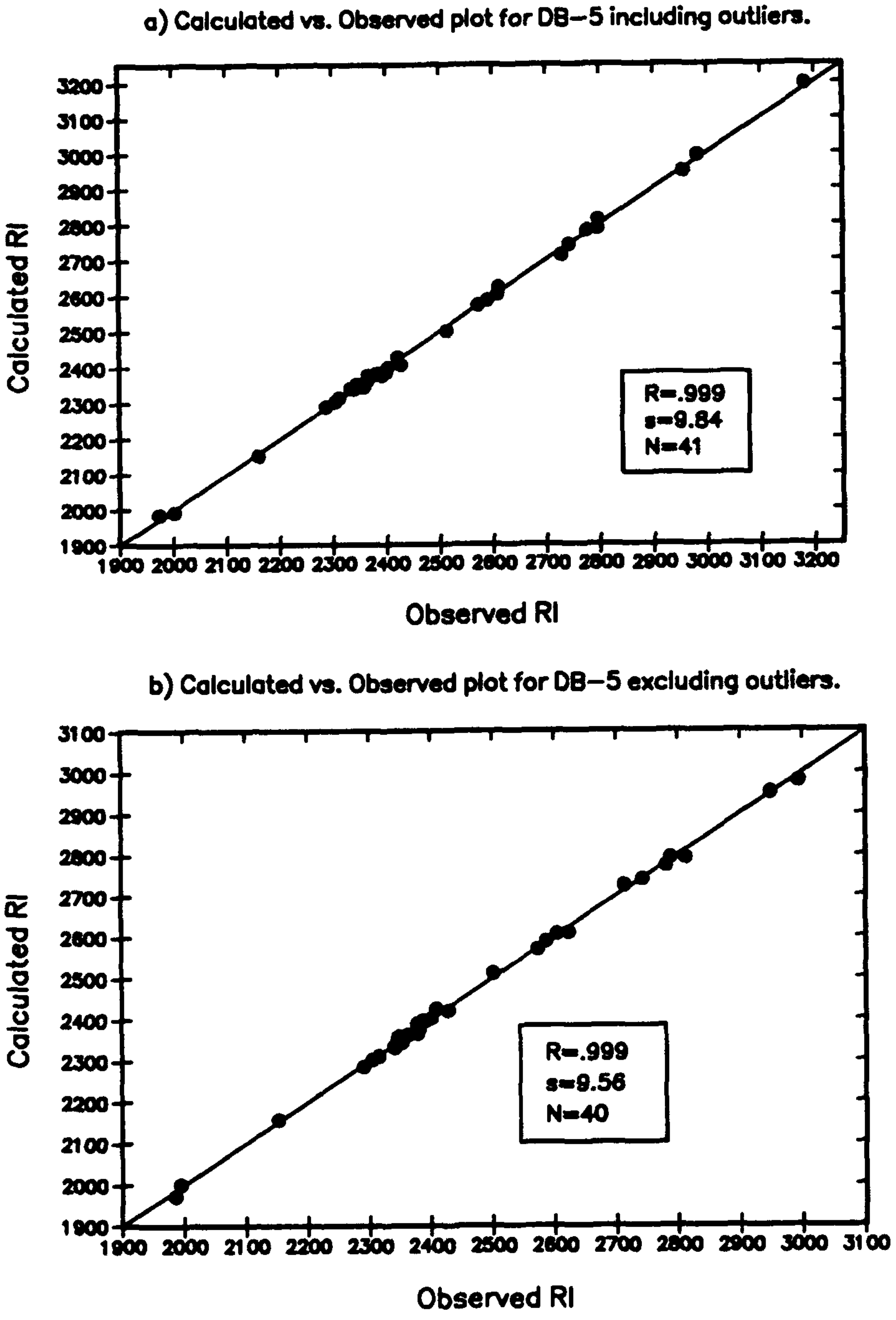
Isomer	DB-5	SE-54	OV-1701	SP-2330	SP-2100
1-monoCl					X
2-monoCl					X
1234-teCl					
1267-teCl				X	
1269-teCl				X	
1278-teCl				X	
1289-teCl				X	
1469-teCl				X	
1478-teCl				X	
2378-teCl				X	X
12378-peCl		X			
1234678-hpCl		X			
12346789-ocCl	X				X

The DB-5 column had the same one outlier with DDG and RRA, and graphically it is obvious that the point has the highest retention time and could exhibit a large leverage on the regression line. Graphs with and without the outlier are shown in Figure 3.8. Since both methods considered the octachloro isomer an outlier and the point did have a large leverage value, the point was dropped from the model.

The SE-54 column showed one DDG outlier and two RRA outliers. Both methods chose the same heptachloro isomer and RRA also chose the pentachloro isomer. A model without the two RRA outliers was chosen, but since the number of observations was below 15, a model with no more than two descriptors could be used.

The OV-1701 column showed two outliers for DDG and none for RRA. Graphically the outliers did not seem to effect the model and DDG values were not large enough to consider these two compounds as outliers. Therefore, the original model and coefficients for N=15 observations is the final model.

The stationary phase SP-2330 showed outliers for both the DDG and RRA methods. While DDG revealed the presence of three outliers, RRA revealed a total of seven outliers. The 1,2,8,9 and the 1,4,6,9 isomers were common to both so they were taken to be true outliers. The other DDG outlier, 1,2,3,4, was not considered an outlier since it seemed to have little impact on the regression line. The remaining RRA outliers were removed from the modeling process as they did tend to influence the regression line. The DDG values for these remaining outliers, although not above the cut off values for three of the five tests, were high enough to make them outliers. The number of observations now changes to 15 and only three descriptors may be used in the final model. Looking at the original data this column had



a) Calculated vs. Observed plot for DB-5 including outliers.

b) Calculated vs. Observed plot for DB-5 excluding outliers.

Figure 3.8 Calculated vs. observed plots for DB-5 before outlier removal (a) and after outlier removal (b).

observations only for the tetrachloro isomers so in reality this model was only a curve fitting exercise. With only the tetrachloro isomers available any predic:·ons outside this area may not be valid. This will be discussed again later.

The final column, SP-2100, showed three DDG outliers and four RRA outliers. The two mono-substituted isomers and the 2,3,7,8 isomer were common to both methods. The mono-substituted isomers had large leverage values as did the octachloro isomer. Since these points could greatly influence the regression line because of the positions on it, they were removed from the model.

A summary of the outliers is shown in Table 3.7. This shows that most of the outliers were found by both methods and that the RRA results were used most of the time. This was done mainly because the RRA method was more robust and therefore more confidence was generated from the RRA results. This was not always the case as is evident in Chapter 4 of this thesis which models dibenzofurans. From the outlier summary three compounds were specifically found to be outliers 3 or more times. These were the 2,3,7,8, the 1,2,3,4,6,7,8, and the 1,2,3,4,6,7,8,9 isomers. The octachloro and heptachloro isomers had large retention times which puts them on the high end of the regression line and therefore they could exert a great deal of leverage. The 2,3,7,8 tetrachloro isomer was an outlier in four different models. This was probably due to its shape as it is the longest molecule and has the greatest L/B ratio. The remaining outliers could be caused by experimental error or a combination of steric effects as well as electronic effects which interact with each column differently to cause the compound not to behave ideally. The exact reasons are not known. New models for all columns, except OV-1701 which had no outliers, are shown in Table 3.8.

Table 3.7 Outlier summary.

Isomer	Frequency as an outlier		
	RRA	DDG	TOTAL
1-Cl	1	1	2
2-Cl	1	1	2
1234-teCl		1	1
1267-teCl	1		1
1269-teCl	1		1
1278-teCl	1		1
1289-teCl	1	1	2
1469-teCl	1	1	2
1478-teCl	1		1
2378-teCl	2	2	4
12378-peCl	1		1
1234678-hpCl	1	2	3
12346789-ocCl	2	1	3

Table 3.8 Summary of final models.

			Column type		
Descriptor	DB-5	SE-54	OV-1701	SP-2330	SP-2100
AVHC					-0.682±0.128
$^7\chi_{CH}$	1350±143			1.945±0.279	
KAPA 3	-430±27.3				-0.345±0.030
KAPA 6				-0.867±0.052	
QNEG				-18.02±2.02	
TOAC 3	972±9.89	0.683±0.025			
EDMN					
ETOT			2.586±0.110		3.766±0.051
L/B	99.4±11.2		0.184±0.034		0.138±0.014
CXVD			-0.335±0.077		
SYMM 35		-0.084±0.018			
CONST.	1074	0.144	-45.16	-1.421	-64.91
R	0.999	0.993	0.996	0.988	0.999
s	9.56(0.4%)	0.016(1.2%)	0.014(1.1%)	0.0088(1.0%)	0.011(0.7%)
F	5862	372	473	152	4197
N	40	13	15	15	35

Model Validation

To validate these models the jackknifed residuals were calculated for each of the final models as shown in Table 3.9. Jackknifing is an excellent way to test the internal validity of a model as discussed in Chapter 2. From the final models calculated versus observed plots and residual plots were obtained and examined to ensure linearity and constancy of variance. Figures 3.9 through 3.13 show the calculated versus observed plots for all the final models. These plots were all reasonable and helped to prove model validity. Again because of the small size of the experimental data sets only internal validation experiments were performed.

Predictions

Since experimental retention times were not available for all classes of isomers (mono-, di-, tri-) on every column, predictions could not be made for all 75 isomers on all columns. It was felt in order to predict the trichloro-dibenzodioxin retention times that at least one trichloro isomer should have been included in the modeling procedure. Since this was not possible, predictions were made only where experimental data was available. For instance, since there were no mono-substituted observations for DB-5, no predictions were made as to their retention index; however, all the other 73 isomers, including the outlying octachloro isomer, were predicted. Predictions of the outliers were made since the experimental value was already present, and a comparison of the predicted versus the observed values could be made. The SP-2330 column had only tetra-substituted observations so the model,

Table 3.9 Jackknifing results.

	DB-5		
Isomer	Observed	JK Estimate	JK Residual
23-diCl	1993	2004	-11
27-diCl	1985	1967	18
28-diCl	1985	1970	15
124-trCl	2152	2159	-7
1234-teCl	2379	2383	-4
1236-teCl	2378	2392	-14
1237-teCl	2382	2382	0
1238-teCl	2382	2378	4
1239-teCl	2392	2398	-6
1246-teCl	2346	2357	-11
1247-teCl	2340	2336	4
1248-teCl	2340	2338	2
1249-teCl	2346	2343	3
1267-teCl	2408	2431	-23
1268-teCl	2349	2360	-11
1269-teCl	2378	2364	14
1278-teCl	2400	2403	-3
1279-teCl	2364	2364	0
1289-teCl	2428	2419	9
1368-teCl	2290	2285	5
1369-teCl	2315	2311	4
1378-teCl	2340	2335	5
1379-teCl	2304	2303	1
1469-teCl	2341	2330	11
1478-teCl	2353	2344	9
2378-teCl	2386	2400	-14

Table 3.9 (Cont.)

Isomer	DB-5 Observed	JK Estimate	JK Residual
12347-peCl	2573	2570	3
12367-peCl	2604	2608	-4
12378-peCl	2587	2590	-3
12389-peCl	2623	2609	16
12468-peCl	2501	2512	-11
12479-peCl	2501	2513	-12
123467-heCl	2812	2791	21
123468-heCl	2742	2739	3
123478-heCl	2781	2772	9
123678-heCl	2788	2794	-6
124679-heCl	2713	2727	-14
124689-heCl	2713	2727	-14
1234678-hpCl	2994	2975	19
1234679-hpCl	2949	2951	-2

Table 3.9 (Cont.)

SE-54

Isomer	Observed	JK Estimate	JK Residual
1368-teCl	1.07500	1.08416	-0.00916
1379-teCl	1.08200	1.07865	0.00335
2378-teCl	1.12500	1.14339	-0.01839
12368-peCl	1.21500	1.19962	0.01538
12379-peCl	1.22500	1.20399	0.02101
12468-peCl	1.19200	1.19176	0.00024
12478-peCl	1.22000	1.21770	0.00230
12479-peCl	1.19200	1.19307	-0.00107
123478-heCl	1.41100	1.41065	0.00035
123678-heCl	1.40900	1.40562	0.00338
123679-heCl	1.33700	1.36185	-0.02485
123689-heCl	1.33700	1.36869	-0.03169
123789-heCl	1.43200	1.39619	0.03581

Table 3.9 (Cont.)

	OV-1701		
Isomer	Observed	JK Estimate	JK Residual
1368-teCl	1.05200	1.04308	0.00892
1379-teCl	1.06300	1.03140	0.03160
2378-teCl	1.10600	1.15517	-0.04917
12368-peCl	1.18900	1.20592	-0.01692
12378-peCl	1.22900	1.21864	0.01036
12379-peCl	1.20300	1.19845	0.00455
12468-peCl	1.17000	1.18491	-0.01491
12478-peCl	1.19600	1.20741	-0.01141
12479-peCl	1.17000	1.15436	0.01564
123478-heCl	1.37000	1.36728	0.00272
123678-heCl	1.36300	1.36407	-0.00107
123679-heCl	1.33800	1.35629	-0.01829
123689-heCl	1.33800	1.35127	-0.01327
123789-heCl	1.39500	1.38808	0.00692
1234678-hpCl	1.58800	1.53194	0.05606

Table 3.9 (Cont.)

	SP-2330		
Isomer	Observed	JK Estimate	JK Residual
1234-teCl	1.01000	1.03333	-0.02333
1236-teCl	1.02000	1.02920	-0.00920
1237-teCl	1.01100	0.99762	0.01338
1238-teCl	1.01100	1.01199	-0.00099
1239-teCl	1.06800	1.07070	-0.00270
1246-teCl	1.00500	1.00126	0.00374
1247-teCl	0.96000	0.94138	0.01862
1248-teCl	0.96000	0.95445	0.00555
1249-teCl	1.00500	1.01471	-0.00971
1268-teCl	0.97700	0.98554	-0.00854
1279-teCl	1.02100	1.01439	0.00661
1368-teCl	0.87600	0.88666	-0.01066
1369-teCl	0.95500	0.94527	0.00973
1378-teCl	0.93500	0.94080	-0.00580
1379-teCl	0.90600	0.91787	-0.01187

Table 3.9 (Cont.)

	SP-2100		
Isomer	Observed	JK Estimate	JK Residual
23-diCl	1.63649	1.63362	0.00287
27-diCl	1.62737	1.60107	0.02629
124-trCl	1.77815	1.77257	0.00558
237-trCl	1.81358	1.81801	-0.00443
1234-teCl	1.99123	1.99744	-0.00622
1236-teCl	1.98900	2.00722	-0.01822
1237-teCl	1.99344	1.99478	-0.00134
1238-teCl	1.99344	1.99959	-0.00615
1239-teCl	2.00432	1.99877	0.00555
1246-teCl	1.95904	1.96375	-0.00471
1247-teCl	1.95279	1.95669	-0.00390
1248-teCl	1.95279	1.94708	0.00571
1249-teCl	1.95904	1.96236	-0.00332
1267-teCl	2.01703	2.04208	-0.02505
1268-teCl	1.96284	1.97859	-0.01575
1269-teCl	1.98767	1.97691	0.01076
1278-teCl	2.01284	2.01955	-0.00671
1279-teCl	1.97818	1.97943	-0.00125
1289-teCl	2.03743	2.03278	0.00464
1368-teCl	1.91009	1.92681	-0.01672
1369-teCl	1.93044	1.92077	0.00967
1378-teCl	1.95665	1 96033	-0.00368
1379-teCl	1.92064	1.91723	0.00341
1469-teCl	1.95231	1.93582	0.01649
1478-teCl	1.96755	1.95706	0.01049
12347-peCl	2.18752	2.18444	0.00308
12378-peCl	2.21219	2.20873	0.00346
12478-peCl	2.16435	2.17102	-0.00667
123478-heCl	2.40483	2.39196	0.01288
123678-heCl	2.42325	2.42114	0.00211
123679-heCl	2.38382	2.38246	0.00136
123789-heCl	2.44091	2.41199	0.02892
124679-heCl	2.34635	2.36239	-0.01604
1234678-hpCl	2.62118	2.61008	0.01109
1234679-hpCl	2.57749	2.59106	-0.01357

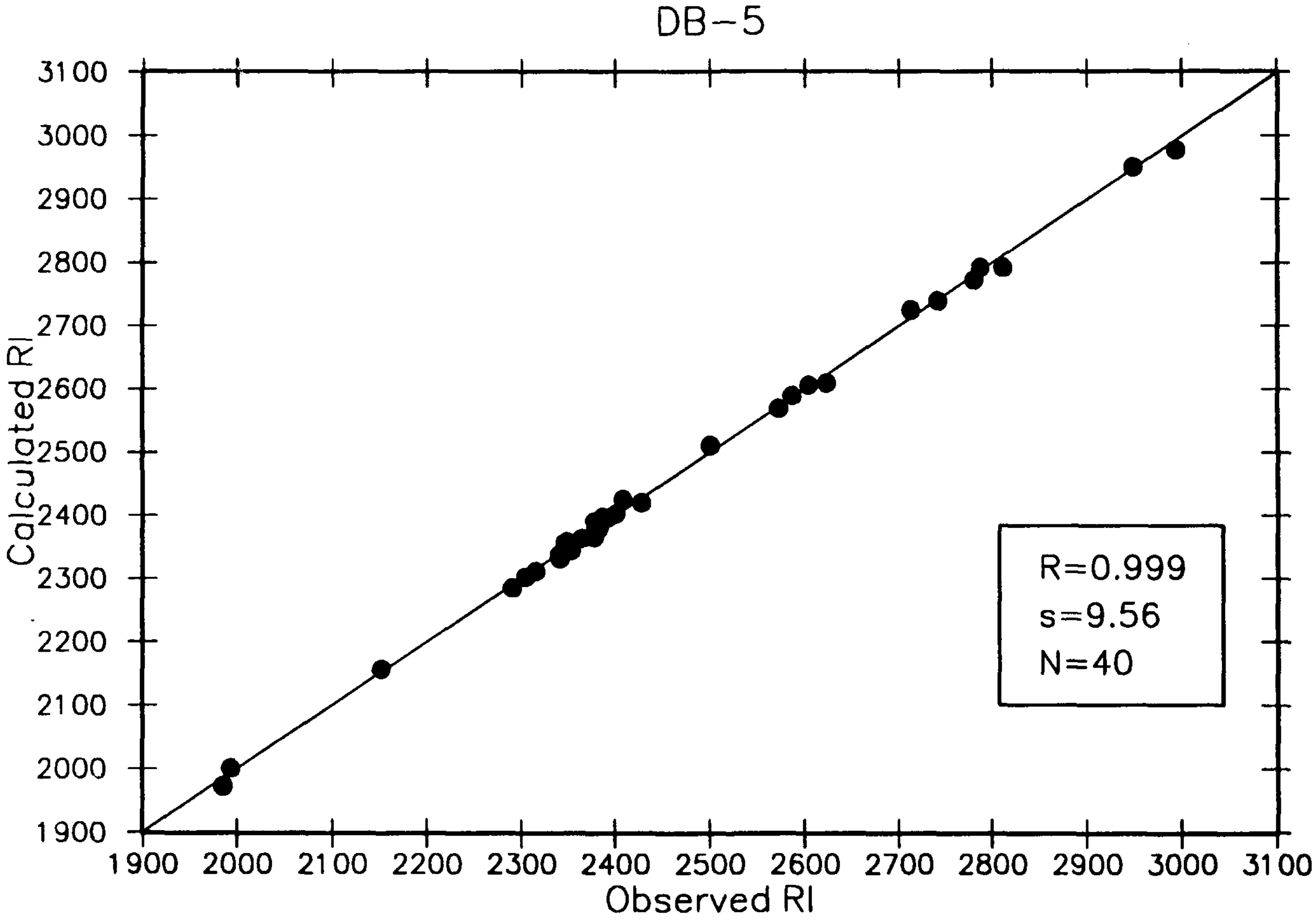


Figure 3.9 Calculated vs. observed plot for the DB-5 column.

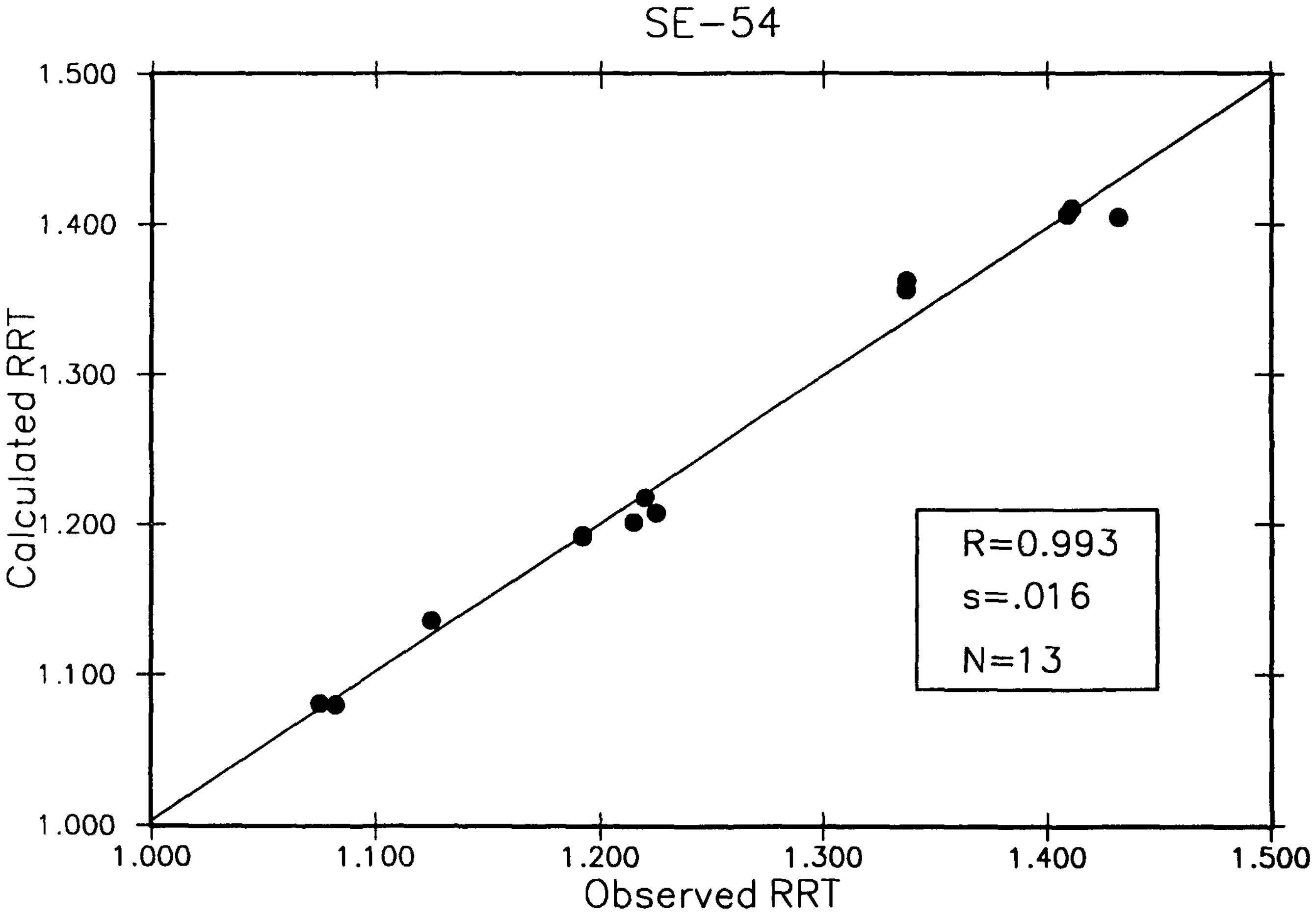


Figure 3.10 Calculated vs. observed plot for the SE-54 column.

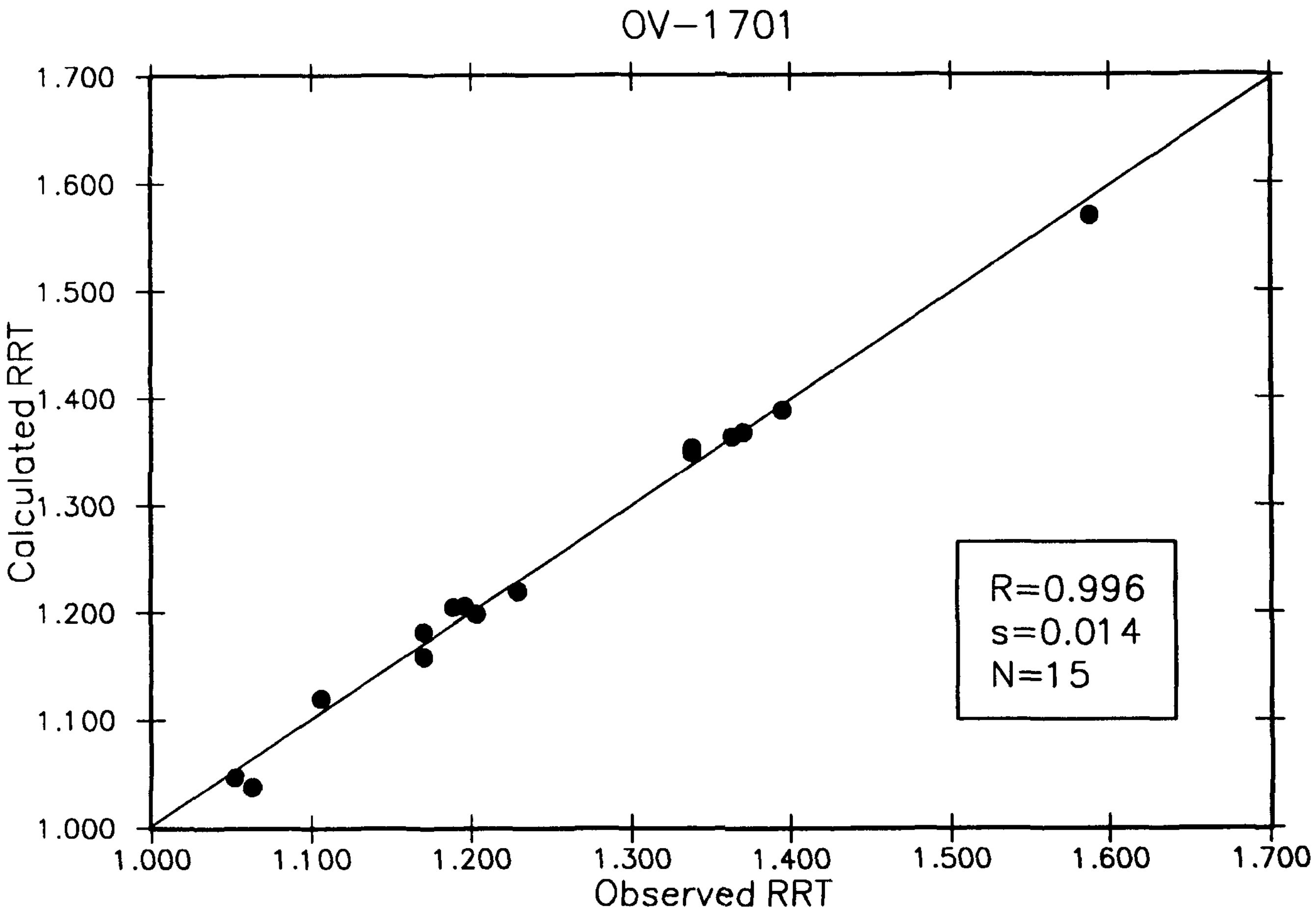


Figure 3.11 Calculated vs. observed plot for the OV-1701 column.

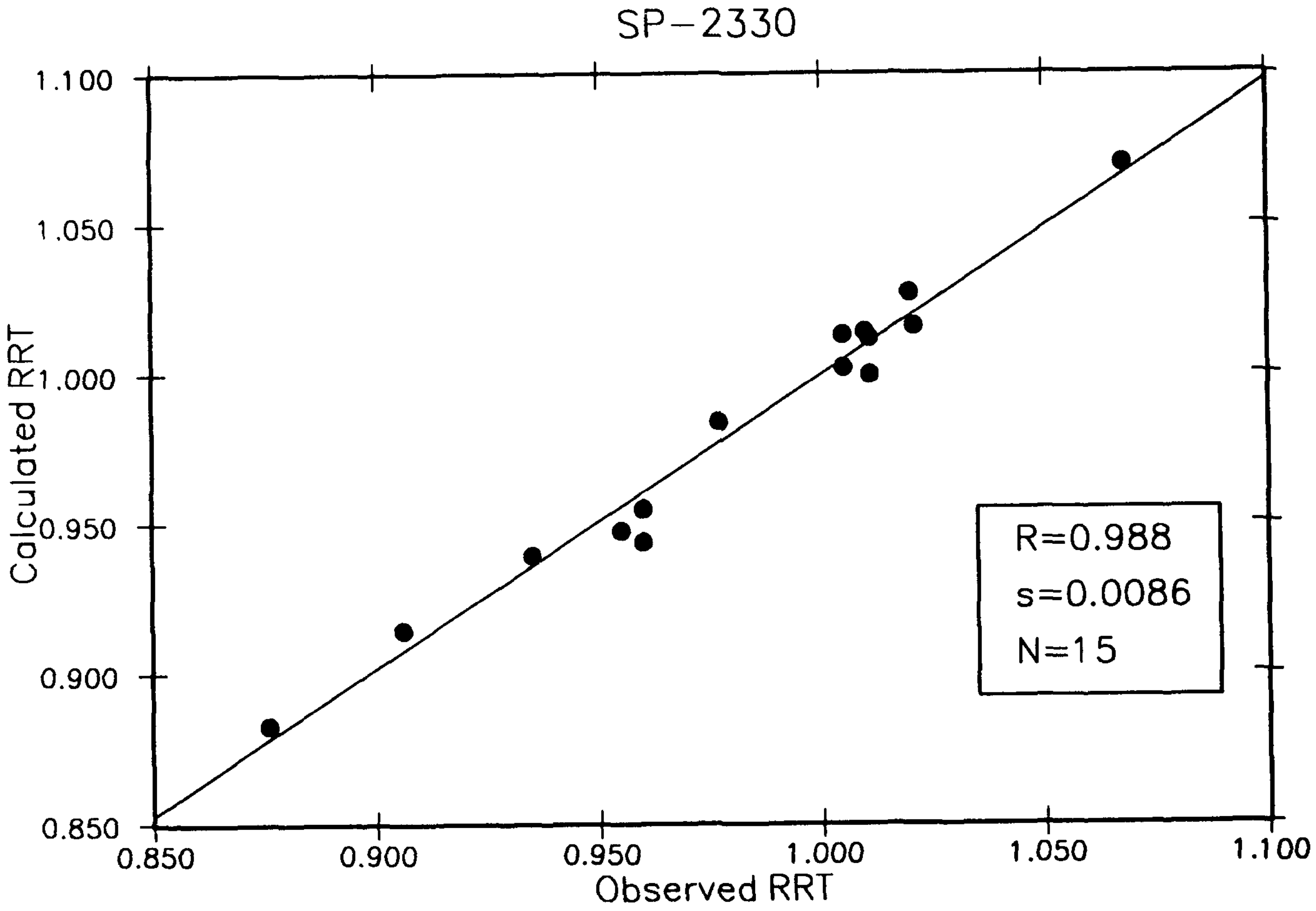


Figure 3.12 Calculated vs. observed plot for the SP-2330 column.

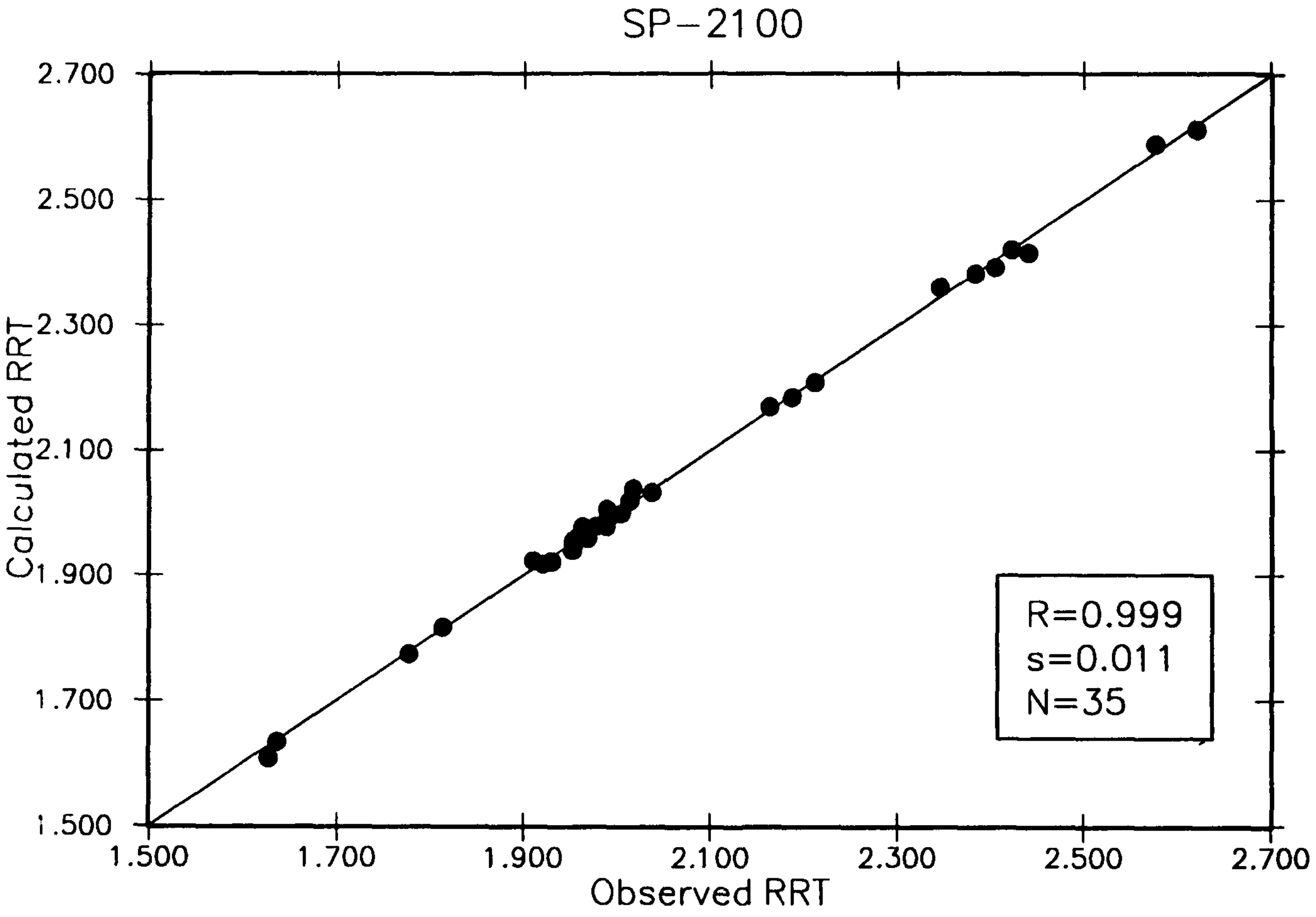


Figure 3.13 Calculated vs. observed plot for the SP-2100 column.

in fact, became a simple curve fitting exercise. Predictions were made for all 22 tetrachloro isomers. Predictions for the other columns were calculated in the same manner and the results for all columns are shown in Table 3.10 along with the errors for the observations used in the model.

As a further test of model validity, the predictions for 73 isomers (excluding mono-substituted) were correlated against each other for the DB-5 and SP-2100 columns. These two columns were picked for two reasons: 1) they had predictions available for nearly all the isomers and 2) the polarities are nearly identical. The correlation coeficient was R=0.9995 or a nearly perfect correlation. This was extremely significant since both models were found independently and do not contain the same descriptors. Only two of the four descriptors were common to both models. This shows that both models were describing the same type of retention behavior and, except for using different experimental conditions, the data were highly correlated. This was the only example where a comparison of columns of nearly identical polarity could be made.

Conclusions

Retention behavior was successfully modeled for most of the 75 isomers of the polychlorinated dibenzodioxins on five different stationary phases of varying polarity. Topological, electronic and geometrical descriptors were used as well as the atom-based descriptors which were new to this type of study. The atom-based descriptors were very important and helped to encode the structural environment of the bridgehead carbons. The models were statistically significant and fit the data extremely well. The usefulness of transformations was realized and increased the fit

Table 3.10 Predicted values for all columns.

	DB-5		
Isomer	Observed	Predicted	Residual
12-diCl		2009	
13-diCl		1941	
14-diCl		1953	
16-diCl		1995	
17-diCl		1975	
18-diCl		1970	
19-diCl		1992	
23-diCl	1993	2001	+ 8
27-diCl	1985	1972	- 7
28-diCl	1985	1973	- 12
123-trCl		2195	
124-trCl	2152	2157	+ 5
126-trCl		2211	
127-trCl		2200	
128-trCl		2191	
129-trCl		2206	
136-trCl		2146	
137-trCl		2132	
138-trCl		2133	
139-trCl		2153	
146-trCl		2159	
147-trCl		2140	
178-trCl		2190	
237-trCl		2192	
1234-teCl	2379	2382	+ 3
1236-teCl	2378	2391	+ 13
1237-teCl	2382	2382	0
1238-teCl	2382	2379	- 3
1239-teCl	2392	2397	+ 5
1246-teCl	2346	2356	+ 10
1247-teCl	2340	2337	- 3
1248-teCl	2340	2338	- 2
1249-teCl	2346	2343	- 3
1267-teCl	2408	2426	+ 18
1268-teCl	2349	2359	+ 10
1269-teCl	2378	2366	- 12

Table 3.10 (Cont.)

	DB-5		
Isomer	Observed	Predicted	Residual
1278-teCl	2400	2403	+ 3
1279-teCl	2364	2364	0
1289-teCl	2428	2421	- 7
1368-teCl	2290	2286	- 4
1369-teCl	2315	2312	- 3
1378-teCl	2340	2335	- 5
1379-teCl	2304	2303	- 1
1469-teCl	2341	2332	- 9
1478-teCl	2353	2345	- 8
2378-teCl	2386	2398	+ 13
12346-peCl	2585	2585	0
12347-peCl	2573	2570	- 3
12367-peCl	2604	2607	+ 3
12368-peCl		2546	
12369-peCl	2561	2561	0
12378-peCl	2587	2590	+ 3
12379-peCl		2563	
12389-peCl	2623	2610	- 13
12467-peCl	2570	2570	0
12468-peCl	2501	2511	+ 10
12469-peCl	2533	2533	0
12478-peCl		2551	
12479-peCl	2501	2512	+ 11
12489-peCl	2560	2560	0
123467-heCl	2812	2793	- 19
123468-heCl	2742	2740	- 2
123469-heCl	2752	2752	0
123478-heCl	2781	2773	- 8
123678-heCl	2788	2793	+ 5
123679-heCl		2755	
123689-heCl		2752	
123789-heCl		2799	
124679-heCl	2713	2725	+ 12
124689-heCl	2713	2726	+ 13
1234678-hpCl	2994	2978	- 16
1234679-hpCl	2949	2951	+ 2
12346789-ocCl	3196	3176	- 20

Table 3.10 (Cont.)

Isomer	SE-54 Observed	Predicted	Residual
1234-teCl		1.118	
1236-teCl		1.066	
1237-teCl		1.067	
1238-teCl		1.067	
1239-teCl		1.064	
1246-teCl		1.051	
1247-teCl		1.059	
1248-teCl		1.059	
1249-teCl		1.045	
1267-teCl		1.100	
1268-teCl		1.051	
1269-teCl		1.049	
1278-teCl		1.080	
1279-teCl		1.049	
1289-teCl		1.094	
1368-teCl	1.075	1.081	+0.006
1369-teCl		1.041	
1378-teCl		1.062	
1379-teCl	1.082	1.080	-0.002
1469-teCl		1.099	
1478-teCl		1.096	
2378-teCl	1.125	1.136	+0.011
12346-peCl		1.216	
12347-peCl		1.229	
12367-peCl		1.209	

Table 3.10 (Cont.)

	SE-54		
Isomer	Observed	Predicted	Residual
12368-peCl	1.215	1.202	-0.013
12369-peCl		1.201	
12378-peCl	1.253	1.223	-0.030
12379-peCl	1.225	1.208	-0.017
12389-peCl		1.209	
12467-peCl		1.198	
12468-peCl	1.192	1.192	0.000
12469-peCl		1.200	
12478-peCl	1.220	1.218	-0.002
12479-peCl	1.192	1.193	+0.001
12489-peCl		1.195	
123467-heCl		1.364	
123468-heCl		1.360	
123469-heCl		1.393	
123478-heCl	1.411	1.411	0.000
123678-heCl	1.409	1.407	-0.002
123679-heCl	1.337	1.357	+0.020
123689-heCl	1.337	1.363	+0.026
123789-heCl	1.432	1.405	-0.027
124679-heCl		1.385	
124689-heCl		1.381	
1234678-hpCl	1.659	1.515	-0.144
1234679-hpCl		1.515	

Table 3.10 (Cont.)

	OV-1701		
Isomer	Observed	Predicted	Residual
1234-teCl		0.959	
1236-teCl		1.010	
1237-teCl		1.031	
1238-teCl		1.039	
1239-teCl		0.995	
1246-teCl		0.986	
1247-teCl		1.017	
1248-teCl		1.006	
1249-teCl		0.983	
1267-teCl		1.042	
1268-teCl		1.024	
1269-teCl		0.976	
1278-teCl		1.019	
1279-teCl		1.035	
1289-teCl		1.021	
1368-teCl	1.052	1.047	-0.005
1369-teCl		0.981	
1378-teCl		1.062	
1379-teCl	1.063	1.038	-0.025
1469-teCl		0.956	
1478-teCl		1.006	
2378-teCl	1.106	1.120	+0.014
12346-peCl		1.176	
12347-peCl		1.181	
12367-peCl		1.211	

Table 3.10 (Cont.)

	OV-1701		
Isomer	Observed	Predicted	Residual
12368-peCl	1.189	1.205	+0.016
12369-peCl		1.169	
12378-peCl	1.229	1.220	-0.009
12379-peCl	1.203	1.199	-0.004
12389-peCl		1.242	
12467-peCl		1.189	
12468-peCl	1.170	1.182	+0.012
12469-peCl		1.148	
12478-peCl	1.196	1.206	+0.010
12479-peCl	1.170	1.159	-0.011
12489-peCl		1.179	
123467-heCl		1.386	
123468-heCl		1.375	
123469-heCl		1.332	
123478-heCl	1.370	1.368	-0.002
123678-heCl	1.363	1.364	+0.001
123679-heC:	1.338	1.354	+0.016
123689-heCl	1.338	1.349	+0.011
123789-heCl	1.395	1.389	-0.006
124679-heCl		1.353	
124689-heCl		1.337	
1234678-hpCl	1.588	1.570	-0.018
1234679-hpCl		1.542	

Table 3.10 (Cont.)

	SP-2330		
Isomer	Observed	Predicted	Residual
1234-teCl	1.010	1.014	+0.004
1236-teCl	1.020	1.027	+0.007
1237-teCl	1.011	1.000	-0.011
1238-teCl	1.011	1.012	+0.001
1239-teCl	1.068	1.070	+0.002
1246-teCl	1.005	1.002	-0.003
1247-teCl	0.960	0.944	-0.016
1248-teCl	0.960	0.955	-0.005
1249-teCl	1.005	1.013	+0.008
1267-teCl	1.100	1.056	-0.044
1268-teCl	0.977	0.984	+0.007
1269-teCl	1.077	1.042	-0.035
1278-teCl	1.054	1.041	-0.013
1279-teCl	1.021	1.016	-0.005
1289-teCl	1.173	1.111	-0.062
1368-teCl	0.876	0.883	+0.007
1369-teCl	0.955	0.948	-0.007
1378-teCl	0.935	0.939	+0.004
1379-teCl	0.906	0.915	+0.009
1469-teCl	1.053	0.975	-0.078
1478-teCl	0.994	0.972	-0.022
2378-teCl	1.000	0.964	-0.036

Table 3.10 (Cont.)

	SP-2100		
Isomer	Observed	Predicted	Residual
1-Cl	0.293	0.257	-0.036
2-Cl	0.299	0.270	-0.029
12-diCl		0.423	
13-diCl		0.370	
14-diCl		0.366	
16-diCl		0.397	
17-diCl		0.385	
18-diCl		0.387	
19-diCl		0.383	
23-diCl	0.433	0.431	-0.002
27-diCl	0.424	0.405	-0.019
28-diCl		0.403	
123-trCl		0.657	
124-trCl	0.600	0.594	-0.006
126-trCl		0.663	
127-trCl		0.648	
128-trCl		0.646	
129-trCl		0.648	
136-trCl		0.580	
137-trCl		0.564	
138-trCl		0.572	
139-trCl		0.570	
146-trCl		0.574	
147-trCl		0.564	
178-trCl		0.635	
237-trCl	0.651	0.656	+0.005
1234-teCl	0.980	0.990	+0.010
1236-teCl	0.975	1.014	+0.039
1237-teCl	0.985	0.988	+0.003
1238-teCl	0.985	0.998	+0.013
1239-teCl	1.010	0.998	-0.012
1246-teCl	0.910	0.919	+0.009
1247-teCl	0.897	0.905	+0.008
1248-teCl	0.897	0.886	-0.011
1249-teCl	0.910	0.916	+0.006
1267-teCl	1.040	1.093	+0.053
1268-teCl	0.918	0.951	+0.033
1269-teCl	0.972	0.951	-0.021

Table 3.10 (Cont.)

	SP-2100		
Isomer	Observed	Predicted	Residual
1278-teCl	1.030	1.044	+0.014
1279-teCl	0.951	0.954	+0.003
1289-teCl	1.090	1.080	-0.010
1368-teCl	0.813	0.838	+0.025
1369-teCl	0.852	0.836	-0.016
1378-teCl	0.905	0.912	+0.007
1379-teCl	0.833	0.829	-0.004
1469-teCl	0.896	0.872	-0.024
1478-teCl	0.928	0.911	-0.017
2378-teCl	1.000	1.055	+0.055
12346-peCl		1.564	
12347-peCl	1.540	1.530	-0.010
12367-peCl		1.687	
12368-peCl		1.502	
12369-peCl		1.508	
12378-peCl	1.630	1.618	-0.022
12379-peCl		1.491	
12389-peCl		1.687	
12467-peCl		1.550	
12468-peCl		1.360	
12469-peCl		1.426	
12478-peCl	1.460	1.481	+0.021
12479-peCl		1.371	
12489-peCl		1.536	
123467-heCl		2.603	
123468-heCl		2.285	
123469-heCl		2.381	
123478-heCl	2.540	2.472	-0.068
123678-heCl	2.650	2.639	-0.011
123679-heCl	2.420	2.413	-0.007
123689-heCl		2.392	
123789-heCl	2.760	2.604	-0.156
124679-heCl	2.220	2.296	+0.076
124689-heCl		2.293	
1234678-hpCl	4.180	4.096	-0.084
1234679-hpCl	3.780	3.878	+0.098
12346789-ocCl	6.760	6.444	-0.316

or significance of the SP-2100 model. Outlier detection was accomplished by two completely different methods and the analysis was carried out in order to obtain the best models for each column. Predictions were made only where valid experimental data existed. Correlations of predictions between similar columns proved to be a valuable tool in validating some of the models as did the jackknifing results. This study indicates the power of QSRR research. These compounds are extremely toxic and studies of this type allow retention data to be predicted instead of experimentally determined. The methods used in this chapter were further refined in the next chapter in which models were obtained for another homogenous data set.

References

(1) Hites, R.A. *Acct. Chem. Res.* **1990,** *23*, 194.

(2) Karasek, F.; Onuska, F. *Anal. Chem.* **1982,** *54*, 309A.

(3) Lawrence,J.; Onuska, F.; Wilkinson, R.; Afghan, B.K. *Chemosphere.* **1986,** *15*, 1085-1090.

(4) Onuska, F.I.; Wilkinson, R.J.; Terry, K. *J. High Res. Chrom. Chrom. Comm.* **1988,** *11*, 9-12.

(5) Lamparski, L.L.; Nestrick, T.J. *Anal. Chem.* **1980,** *52*, 2045-2054.

(6) Hollaway, T.T.; Fairless, B.J.; Freidline, C.E.; Kimball, H.E.; Kloepfer, R.D.; Wurrey, C.J.; Jonooby, L.A.; Palmer, H.G. *Appl. Spect.* **1988,** *42*, 359-369.

(7) Kováts, E. *Chimia.* **1968,** 459.

(8) Van den Dool, H.; Kratz, P. Dec. *J. Chrom.* **1963,** *11*, 463.

(9) O'Keefe, P.W.; Smith, R.; Meyer, C.; Hilker, D.; Aldous, K.; Jelus-Tyror, B. *J. Chrom.* **1982,** *242*, 305-312.

(10) Donnelly, J.R.; Munslow, W.D.; Mitchum, R.K.; Sovocool, G.W. *J. Chrom.* **1987,** *392*, 51-63.

(11) Humppi, T.; Heinola, K. *J. Chrom.* **1985,** *331*, 410-418.

(12) Oehme, M.; Kirschner, P. *Anal. Chem.* **1984,** *56*, 2754-2759.

(13\ Korfmacher, W.A.; Mitchum, R.K. *J. High Res. Chrom. Chrom. Comm.* **1982,** 681-682.

(14) Allinger, N.L.; Yul, Y.H.; MM2/MMP2, 85-Force Field (QCPE Program No. 395). Indiana University, IN: Quantum Chemistry Program Exchange, 1985.

(15) Burkert, U; Allinger, N.L. *Molecular Mechanics*; ACS Monograph 177; American Chemical Society: Washington, DC, 1982.

(16) MOPAC, ver 5.0. Quantum Chemistry Program Exchange, QCPE Program No. 445.

(17) Bermejo, J.; Guillén, M. *Anal. Chem.* **1987,** *59*, 94-97.

(18) Raymer, J.; Weisler, D.; Novotny, M. *J. Chrom.* **1985,** *325*, 13-22.

(19) Kier, L.B.; Hall, L.H. *Molecular Connectivity in Structure Activity Relationships*; John Wiley & Sons, Inc: New York, 1986.

(20) Kier, L.B. *Quant. Struct.-Act. Relat. Pharmocol., Chem. Biol.* **1985**, *4(3)*, 109.

(21) Abraham, R.J.; Griffiths, L.; Loftus, P. *J. Comput. Chem.* **1982**, *3*, 407-416.

(22) Abraham, R.J.; Smith, P.E. *J. Comput. Chem.* **1988**, *9*, 288-297.

(23) Dixon, S.L.; Jurs, P.C. *Empirical Calculations of Partial Atomic Charges in Organic and Ionic Compounds*, in preparation.

(24) Yates, K. *Hückel Molecular Orbital Theory*; Academic: New York. 1980.

(25) Stanton, D.T.; Jurs, P.C. *Anal. Chem.* **1990**, *62*, 2323.

(26) Anker, L.S.; Edwards, P.A.; Jurs, P.C. *Anal. Chem.* **1990**, *62*, 2676-2684.

(27) Radecki, A.; Lamparczyk, H.; Kaliszan, R. *Chromatographia.* **1979**, *12*, 58.

(28) Lowe, J.P. *Quantum Chemistry*; Academic: New York, 1978.

(29) Robbat, A. Jr.; Kalogeropoulos, C. *Anal. Chem.* **1990**, *62*, 2684-2688.

(30) Rousseeuw, P.J. *J. Am. Stat. Assoc.* **1984**, *79*, 871-880.

(31) Massart, D.L.; Kaufman, P.J.; Rouseeuw, P.J.; Lerox, A. *Anal. Chem. Acta.* **1986**, *187*, 171-179.

Chapter 4

PREDICTION OF GAS CHROMATOGRAPHIC RETENTION DATA FOR POLYCHLORINATED DIBENZOFURANS

Polychlorinated dibenzofurans have also been intensely studied recently. Their toxicity is well known and their presence has been reported in water as well as in fly ash from incinerators (1-3). Trace analysis of the furans is very similar to the dioxins. Gas chromatography has been used to separate many of the PCDF isomers and their retention data in the form of retention indexes (4) and relative retention times (5-8) has been reported on many different stationary phases. The relative retention times are usually relative to the 2,3,7,8 tetrachloro dibenzofuran which is the most toxic (1). Some of the experiments used other compounds, such as anisoles, as the elution standard for relative retention times.

There are 135 polychlorinated dibenzofurans and this is a closed data set. Although similar to the dioxins, there is only one oxygen bridging the space between the two aromatic rings instead of two. The isomers can have chlorines attached to positions 1-4 and/or 6-9 as shown in Figure 4.1. The 135 isomers and their retention data are shown in Table 4.1.

The strength of the interactions between the stationary phase and the compound placed on a chromatographic column determine the retention of the compound. The interactions are based upon electronic, chemical and structural properties of the compounds. The time of retention on a column is usually unique to the compound and forms the basis of separation.

DIBENZOFURAN

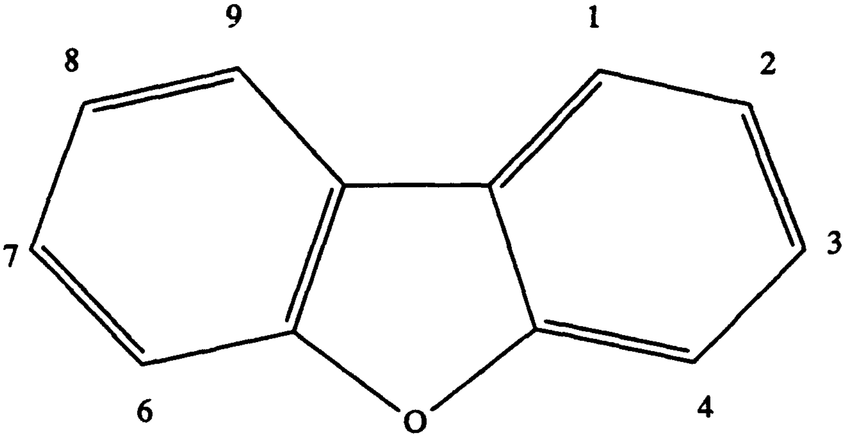


Figure 4.1 The structure and numbering scheme of the polychlorinated dibenzofurans.

Table 4.1 The 135 polychlorinated dibenzofurans and retention data.

	SP-2330	OV-101	DB-5	SE-54	OV-1701	DB-5
Isomers	RRT	RRT	RI	RRT	RRT	RRT
1-Cl			1739			0.341
2-Cl			1749			0.443
3-Cl			1749			0.439
4-Cl			1760			0.457
12-diCl			1934			
13-diCl			1884			
14-diCl			1913			
16-diCl						
17-diCl			1910			
18-diCl			1925			
19-diCl			1975			
23-diCl			1939			
24-diCl			1912			
26-diCl			1946			0.626
27-diCl			1930			0.611
28-diCl			1935			0.615
34-diCl			1959			
36-diCl			1944			
37-diCl			1930			
46-diCl			1953			
123-trCl			2113			
124-trCl			2085			
126-trCl			2125			
127-trCl			2109			
128-trCl			2129			
129-trCl						
134-trCl			2088			
136-trCl			2072			0.748
137-trCl			2057			0.747
138-trCl			2070			0.752
139-trCl			2124			
146-trCl			2094			
147-trCl			2086			
148-trCl			2100			
149-trCl			2151			

Table 4.1 (Cont.)

	SP-2330	OV-101	DB-5	SE-54	OV-1701	DB-5
Isomers	RRT	RRT	RI	RRT	RRT	RRT
234-trCl			2148			0.831
236-trCl			2141			
237-trCl			2134			
238-trCl			2132			0.805
239-trCl			2111			
246-trCl			2101			
247-trCl			2099			
248-trCl			2097			
249-trCl			2082			
346-trCl			2152			
347-trCl			2150			
348-trCl			2151			0.824
349-trCl			2125			
1234-teCl	0.800	0.978	2310			
1236-teCl			2307			
1237-teCl	0.766	0.950	2294			
1238-teCl	0.805	0.967	2307			
1239-teCl			2369			
1246-teCl			2264			
1247-teCl			2264			
1248-teCl			2274			0.949
1249-teCl			2335			
1267-teCl	0.873	0.995	2329			
1268-teCl			2281			
1269-teCl			2364			
1278-teCl	0.840	0.989	2322			
1279-teCl	0.875	1.005	2341			
1289-teCl			2406			
1346-teCl			2262			
1347-teCl			2257			
1348-teCl			2276			
1349-teCl			2325			
1367-teCl	0.713	0.937	2272			
1368-teCl	0.625	0.889	2227			
1369-teCl			2296			

Table 4.1 (Cont.)

	SP-2330	OV-101	DB-5	SE-54	OV-1701	DB-5
Isomers	RRT	RRT	RI	RRT	RRT	RRT
1378-teCl			2263			
1379-teCl	0.687	0.938	2273			
1467-teCl	0.806	0.954	2288			
1468-teCl			2242			
1469-teCl			2314			
1478-teCl			2290			
2346-teCl	1.029	1.006	2339			1.003
2347-teCl	0.970	1.005	2337			
2348-teCl	1.008	1.002	2340			1.000
2349-teCl			2308			
2367-teCl	1.042	1.017	2354			
2368-teCl	0.891	0.959	2297			0.964
2378-teCl	1.000	1.000	2338	1.111	1.114	1.000
2467-teCl	0.934	0.967	2305			
2468-teCl			2254	1.064	1.061	
3467-teCl			2362			
12346-peCl			2496			
12347-peCl			2495			1.153
12348-peCl			2508			
12349-peCl						
12367-peCl	1.078	1.226	2540			
12368-peCl						
12369-peCl			2546			
12378-peCl	1.040	1.217	2507	1.215	1.193	1.154
12379-peCl						
12389-peCl			2593			
12467-peCl			2465			
12468-peCl	0.842	1.100		1.158	1.138	
12469-peCl			2497			
12478-peCl	0.939	1.164		1.190	1.168	1.124
12479-peCl	0.959	1.181	2479			
12489-peCl			2559			
13467-peCl			2469			
13468-peCl						
13469-peCl						

Table 4.1 (Cont.)

	SP-2330	OV-101	DB-5	SE-54	OV-1701	DB-5
Isomers	RRT	RRT	RI	RRT	RRT	RRT
13478-peCl			2469			
13479-peCl			2473			
13489-peCl						
23467-peCl	1.465	1.271	2555			
23468-peCl			2495	1.206	1.206	
23469-peCl			2476			
23478-peCl	1.403	1.258	2551	1.243	1.242	1.193
23479-peCl			2467			
23489-peCl	1.104	1.237	2521			
123467-heCl	1.424	1.540	2706			
123468-heCl			2650	1.318	1.279	
123469-heCl						
123478-heCl	1.370	1.542	2708			
123479-heCl			2720			
123489-heCl						
123678-heCl	1.384	1.554		1.371	1.330	1.326
123679-heCl						
123689-heCl	1.587	1.604				
123789-heCl						
124678-heCl	1.225	1.453		1.324	1.287	1.287
124679-heCl						
124689-heCl	1.401	1.494	2686	1.348	1.311	
134678-heCl	1.199	1.454				
134679-heCl						
234678-heCl	2.001	1.603	2748	1.406	1.400	1.364
1234678-hpCl	1.834	1.998	2898	1.567	1.495	
1234679-hpCl			2913			
1234689-hpCl	2.084	2.061	2922	1.598	1.526	
1234789-hpCl			2986			
12346789-ocCl			3147			1.798

Experimental Section

The available data for the furans included retention indexes as well as relative retention times on five different stationary phases ranging from the polar SP-2330 to the non-polar OV-101. Two different data sets were found for the same column, DB-5, one reported retention indexes (DB-5 RI) and the other reported relative retention times (DB-5 RRT). This gave a total of six data sets to model. The six data sets were experimentally determined from four different sources; two of the sources cited data on two different columns. Column parameters and references are shown in Table 4.2.

One of the data sets, DB-5 RI, reported data for 115 out of the 135 isomers. This was the largest data set. The other sets had 35 isomers for both SP-2330 and OV-101, 14 isomers for SE-54 and OV-1701, and 26 isomers for DB-5 RRT. The experimental error was not reported in most cases. The DB-5 RI experimental error was later found to be approximately seven index units (9). The error for the other data sets was assumed to be about 1% at the mean of the range.

This study was conducted in four stages as described in Chapter 2: 1) Entry, modeling and storage of the structures in three-dimensions and the associated retention data; 2) Generation of descriptors; 3) Multiple linear regression analysis; and 4) Model validation.

Table 4.2 Column parameters.

Manufacturer[a]	J&W	N/A	N/A	Supelco	Shim.	J&W
Column Type[b]	DB-5	SE-54	OV-1701	SP-2330	OV-101	DB-5
Column I.D.(mm)	N/A	0.3	0.3	0.20	0.20	0.25
Column Length (m)	30	25	20	60	50	30
Film Thickness (µm)	N/A	0.10	0.10	N/A	N/A	N/A
Carrier Gas	He	He	He	N_2	N_2	He
Data[c]	RI	RRT	RRT	RRT	RRT	RRT
Temperature	175 (1 min)	60-260	60-260	180 (2 min)	200 (2min)	50-200 (rapid)
Program (°C)	175-300 @6°/min	10°/min	10°/min	180-250	200-250	200-320
				2°/min	2°/min	10°/min
Reference	(4)	(5)	(5)	(6)	(6)	(7)

a) Columns obtained from J&W Scientific, Supelco, or Shimadzu.

b) DB-5 is a methyl silicone column with 5% phenyl substitution.
SE-54 is similar to DB-5. It is being replaced by DB-5.
OV-1701 is 86% dimethyl polysiloxane and 14% cyanopropyl phenyl.
SP-2330 is 90% bis cyanopropyl and 10% phenyl cyanopropyl polysiloxane.
OV-101 is 100% dimethyl polysiloxane.

c) RI=Retention Index .
RRT=Relative Retention Time which is relative to 2,3,7,8 tetrachloro dibenzodioxin which = 1.000 min. except for SE-54 and OV-1701 which are relative to 5-chloro-2-(2,4-dichlorophenoxy)-anisole which =1.000 min. Absolute retention times for the standard on SE-54 and OV-1701 is 18.14 min. and 17.54 min. respectively.

Entry, Storage and Molecular Modeling

The structures of each of the 135 PCDF isomers were entered into a Sun 4/110 workstation using the ADAPT subroutine UDRAW (10) and stored as a connection table as described in Chapter 2. Three-dimensional models were generated with the molecular modeling routine of ADAPT (11). The structure of the furans was nearly planar. This was confirmed using the molecular modeling routines MOPAC (12) and Allinger's MM2 (13,14). This was understandable since their are two aromatic rings with a fairly rigid bridge in between.

Descriptor Generation

The structures of the PCDFs were numerically encoded in the form of descriptors. Four types of descriptors were calculated: topological, electronic, geometrical and atom-based descriptors. The atom-based descriptors encoded information about the environment surrounding the bridgehead carbons.

Topological. The only topological descriptor used was a simple path 1 molecular connectivity index developed by Kier and Hall (15). This type of descriptor can be calculated using the connection table data only. This descriptor has been used in other QSRR studies and correlates well with retention data (16-18).

Electronic. Only one purely electronic descriptor was used. This was the descriptor QSUM which was calculated using equation 1 where Q is the sum of the absolute values of all partial atomic charges (q_i). The partial atomic charges were calculated using the method described by Abraham and Smith (19,20) and later modified for use with ADAPT (21).

$$Q=\Sigma \; |q_i| \qquad (1)$$

Geometrical. These descriptors require a three-dimensional model of the compound stored as x, y and z coordinates. A total of four geometrical descriptors were used. The descriptor SHDW3 is defined as the area of the shadow a molecule projects onto a two-dimensional plane by disregarding the third dimension. The plane for SHDW3 is the Y-Z plane. The first and second moments of inertia were aligned with the X and Y axis and since the molecule was relatively flat, this descriptor encoded the smallest possible shadow area. The largest shadow area would be a projection onto the X-Y plane. The algorithm for this calculation can be found in Stouch and Jurs (22), but was further refined by Rohrbaugh and Jurs (23). All compounds were stored in the same orientation to overcome any problems varying orientations might create. The next descriptor was a three-dimensional Weiner index (3D-W). The Weiner index (24) is the sum of the distances between pairs of atoms using the values from the connection table. The 3-D Weiner index uses throughspace distances determined from the three-dimensional models. The descriptor WPSA2 is a charged partial surface area descriptor based upon the partial atomic charges calculated using the method developed by Abraham and Smith (19,20) but then related to the surface area of the molecule. These descriptors were developed by Stanton (25). WPSA2 was calculated using equation 2 where +SA is the positive surface area for atom i with a positive partial atomic charge and Q_T+ is the total positive charge. The total surface area is the solvent accessible surface area using water as a solvent (26). The 1000 is a scaling factor. The final geometrical

$$\mathrm{WPSA2} = \left\{ [\, \Sigma\,(+\mathrm{SA})\, Q_T{}^{+}\,] \right\} \left\{ \frac{\text{Total Surface Area}}{1000} \right\} \qquad (2)$$

descriptor is MOMH2 which is the second major moment of inertia of the molecule with hydrogens attached. The units of MOMH2 are mass times distance squared (AMU-Angstroms^2) (27).

Atom-Based. At first the atom-based descriptors calculated for the PCDFs were similar to the atom-based descriptors calculated for the PCDDs in that the environment of the four bridgehead carbons was chosen. This lead to many descriptors being calculated for the four carbon center environment. Most of these descriptors were deleted during objective feature selection and initial attempts at building regression models. The atom-based descriptors were not proving to be as valuable as they were with the dioxins. A new atom selection was made to try to develop descriptors more adept at encoding structural information about the bridgehead. Another set of atom-based descriptors were calculated to describe the environment surrounding the two carbons bonded to the one oxygen (Figure 4.2). Therefore, there are two different sets of atom-based descriptors presented in the models and the descriptors calculated are referred to as (4C) or (2C); for example TOAC1 (4C) and TOAC1 (2C) for the four and two active carbon center cases respectively.

A total of six different atom-based descriptors were calculated for the final models. The descriptor CSTR3 was calculated for both the (2C) and (4C) environments. CSTR3 is the sum over all chlorines of the chlorine Van der Waals energy divided by the throughspace distance between the chlorine and the carbon

Atom-Based Descriptors

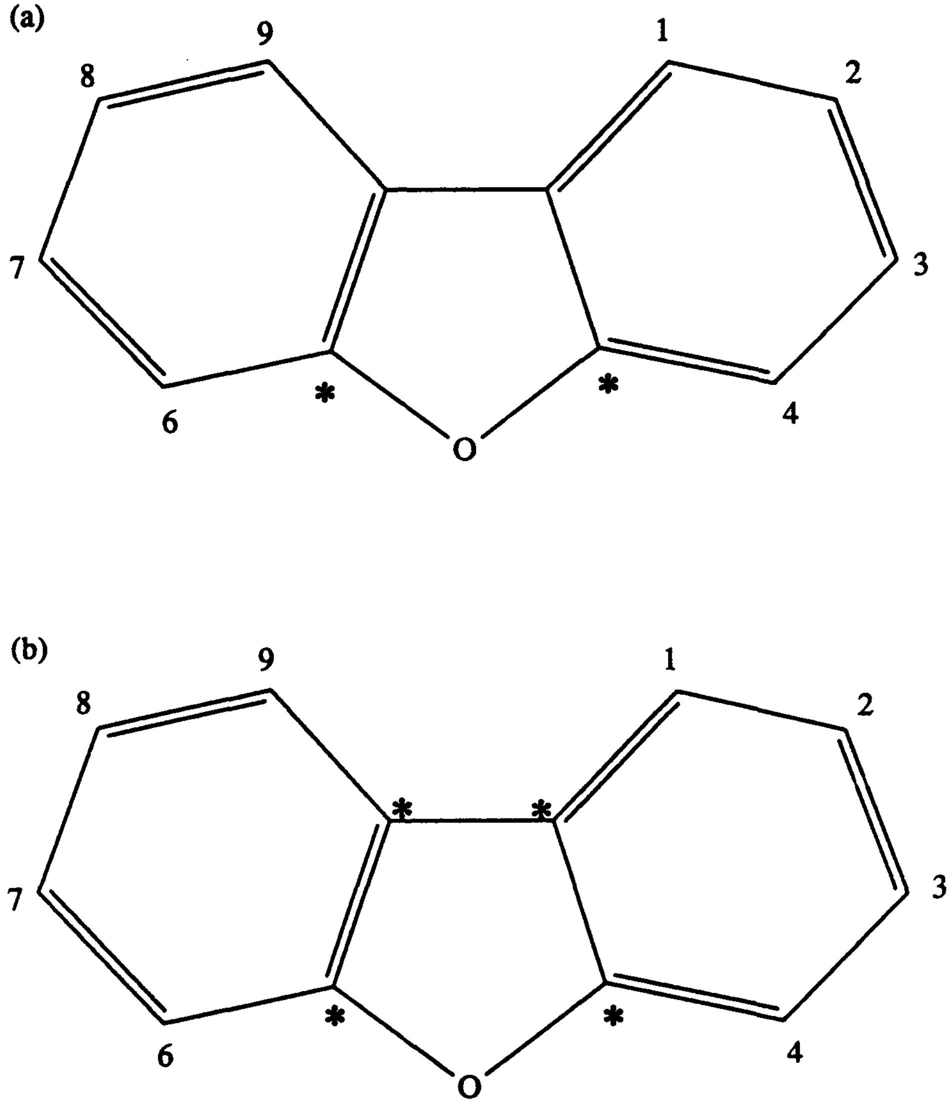


Figure 4.2 Positions of the activated carbon centers of interest for the atom-based descriptors (*). For some descriptors only two carbons were active (a) and for others all four bridgehead carbons were active (b).

center raised to the third power (28). The next descriptor, TOAC1 (2C), is the sum of the absolute values of the partial atomic charges for all heavy atoms (non-hydrogens) one bond from the two carbon centers. TOHC3 (2C) is the sum of the absolute values of the Hückel charges for all heavy atoms three bonds away from the two carbon centers. WHK2-1 (2C) is the sum of the Hückel charges for all heavy atoms two bonds away from the two carbon centers. The Hückel charges were determined from extended Hückel theory (29,30). The final atom-based descriptor was MNAC1 (4C) which is the most negative atomic charge among the heavy atoms one bond away from the four carbon center. The atomic charges were Abraham and Smith charges (19,20).

Six of the final 11 descriptors were atom-based as shown in the descriptor summary in Table 4.3. Atom-based descriptors were able to describe the environment of either the four bridgehead carbons or the two carbons bonded to the oxygen. There were four (2C) descriptors and two (4C) descriptors showing the importance of the area surrounding the oxygen. The actual values of the descriptors are averaged over either the two or four carbon environment to determine a single value.

Regression Analysis

Well over 300 descriptors were calculated for the 135 PCDF isomers. The atom-based descriptors alone accounted for over half of the total descriptors calculated. To delete most of the descriptors which contained redundant information, objective feature selection was performed. The goal of objective feature selection was to eliminate any high pairwise correlations of R>0.90 and

Table 4.3 Descriptor summary.

MOLC1 (a)	Simple path 1 molecular connectivity.
QSUM (b)	Sum of the absolute values of the atomic charges.
SHDW3 (c)	Area of the shadow projected onto the Y-Z plane when the compound has its first and second moments of inertia aligned with the X and Y axis.
3D-W(c)	Three-dimensional Weiner index using throughspace distances.
WPSA2 (c)	Weighted positive charged surface area.
MOMH2 (c)	Second major moment of inertia with hydrogens attached.
CSTR 3 (d) **(2C/4C)**	Sum over all the chlorines of the Van der Waals energy of the chlorine divided by the throughspace distance raised to the third power from the carbon center of interest to the chlorine.
TOAC1(2C)(d)	Sum of the absolutes values of the atomic charges for all heavy atoms 1 bond away and averaged over the two carbon centers.
TOHC3(2C)(d)	Sum of the absolute values of the Hückel charges for all heavy atoms 3 bonds away from the two carbon centers.
WHK2 (2C)(d)	The weighted sum of the Hückel charges on heavy atoms two bonds away from the two carbon centers.
MNAC1(4C)(d)	Most negative atomic charge among heavy atoms 1 bonds away.

(a) Topological
(b) Electronic
(c) Geometrical
(d) Atom-based

eliminate any descriptors containing mostly zeros or a significant number of identical values. These descriptors have little or no information content. Descriptors with very low standard deviations were also eliminated for the same reason. A considerable pool of descriptors remained and were subjected to rigorous multicollinearity testing as described in Chapter 2. The set of descriptors containing the most unique and relevant information were subjected to initial attempts at regression analysis. When initial models showed little promise, new descriptors were calculated. These were the atom-based descriptors which concentrated on the environment surrounding the carbons bonded to the oxygen. These new descriptors were subjected to objective feature selection and tested for multicollinearities along with the other descriptors until a final pool of descriptors remained. This final pool was shown in Table 4.3. The descriptors were then submitted to interactive or forward stepwise regression analysis (31) to determine the optimum model for each of the six data sets.

Results and Discussion

Models were obtained for each of the data sets and contained from two to five descriptors each. Although one of the data sets contained over 100 observations, only four descriptors were needed to model the column successfully. Any more descriptors would, perhaps, overfit the data.

As briefly discussed earlier, new atom-based descriptors were calculated based upon a two carbon environment instead of the four carbon environment. This change reflects the idea from Chapter 3 of determining the environment near the oxygen topologically, electronically and geometrically. For the dioxins there were two oxygens and four carbon centers were used. This was not as productive for the furans, and a switch to the two carbon centers was made in an attempt to refine the oxygen's environment further. As with the dioxins, positioning the chlorines near the oxygen increases the retention time on the column. Any descriptors which could numerically encode this structural feature would be a good candidate for a model. The (2C) descriptors survived objective feature selection and initial regression analysis much better for the furans although two (4C) descriptors did remain throughout the modeling process.

The column with the greatest number of observations was the DB-5 RI column in which Hale et al. reported 115 retention indexes for the 135 isomers (4). Many different models were developed but the best model was a four descriptor model given in equation 3. The model was one of the best models obtained for any

$$\begin{array}{rrl} \mathrm{RI} = & 529.4 & (\mathrm{MOLC1}) + \\ & 11.61 & (\mathrm{WPSA2}) + \\ & -152.4 & (\mathrm{WHK2}) + \\ & 0.056 & (\mathrm{MOMH2}) + \\ & -1764 & (\mathrm{Const.}) \end{array} \qquad (3)$$

R=0.999 s=14.42 (0.6%) F=10062 N=115

column as can be seen with the low error s=14.42 or about 0.6%. The experimental reproducibility was about seven retention index units. The F-statistic shows the high degree of statistical credibility of the model and demonstrates the superb fit of the calculated values

The same procedure was used to generate regression models for the other five data sets. The observation to descriptor ratio never violated the 5:1 rule as described in Chapter 2; therefore, some models could only have two descriptors. The coefficients for the initial regression models for all data sets are shown in Table 4.4.

Initially excellent models were developed for all of the stationary phases. The multiple correlation coefficients and associated errors were acceptable and in most cases, superb. However, upon graphical analysis of the calculated versus observed and residual plots, a non-linearity problem in the OV-101 model became apparent. The calculated versus observed plot shown in Figure 4.3 shows the non-linear problem; however, it is more noticeable in the residual plot shown in Figure 4.4. To solve this problem, graphs of the dependent variable versus the independent variables were generated and analyzed to determine the cause of the non-linearity. Neither of these plots showed any significant problems with non-linearity. The next step would have been a transformation of the independent variable or x-axis if any problems were evident; however, this would not have proved valuable since the residual plot also showed a separate problem, non-constant variance. The non-constant variance is shown in Figure 4.4 as the increase in the error as the dependent variable increases (31). The variance problem is very subtle, but since it was present a transformation of the dependent variable or y-axis was attempted as discussed in Chapter 2. Transformations of the y-axis are not usually attempted first because these transformations will normally effect the error variance. Since the error variance was already a problem, a transformation of the y-axis could solve both the non-linearity problem and the variance problem simultaneously.

Table 4.4 Summary of models.

Descriptors	DB-5RI	DB-5RRT	OV-1701	OV-101	SE-54	SP-2330
MOLC1	529.4	0.449				
QSUM			1.652			
SHDW3						-0.0292
3D-W				0.00402	0.00149	0.00366
WPSA2	11.61			0.0193		
MOMH2	0.0556					
CSTR3(2C)			0.709		0.624	3.442
CSTR3(4C)		0.448				
TOAC1(2C)						7.741
TOHC3(2C)						-0.461
WHK2(2C)	-152.4					
MNAC1(4C)		-5.454				
CONST.	-1764	-2.413	-1.579	-2.180	0.215	-1.746
R	0.999	0.999	0.988	0.990	0.993	0.982
s	14.42(0.6%)	0.015(1.4%)	0.023(1.8%)	0.044(3.0%)	0.019(1.4%)	0.074(5.5%)
F	10062	4310	229	811	425	174
N	115	26	14	35	14	35

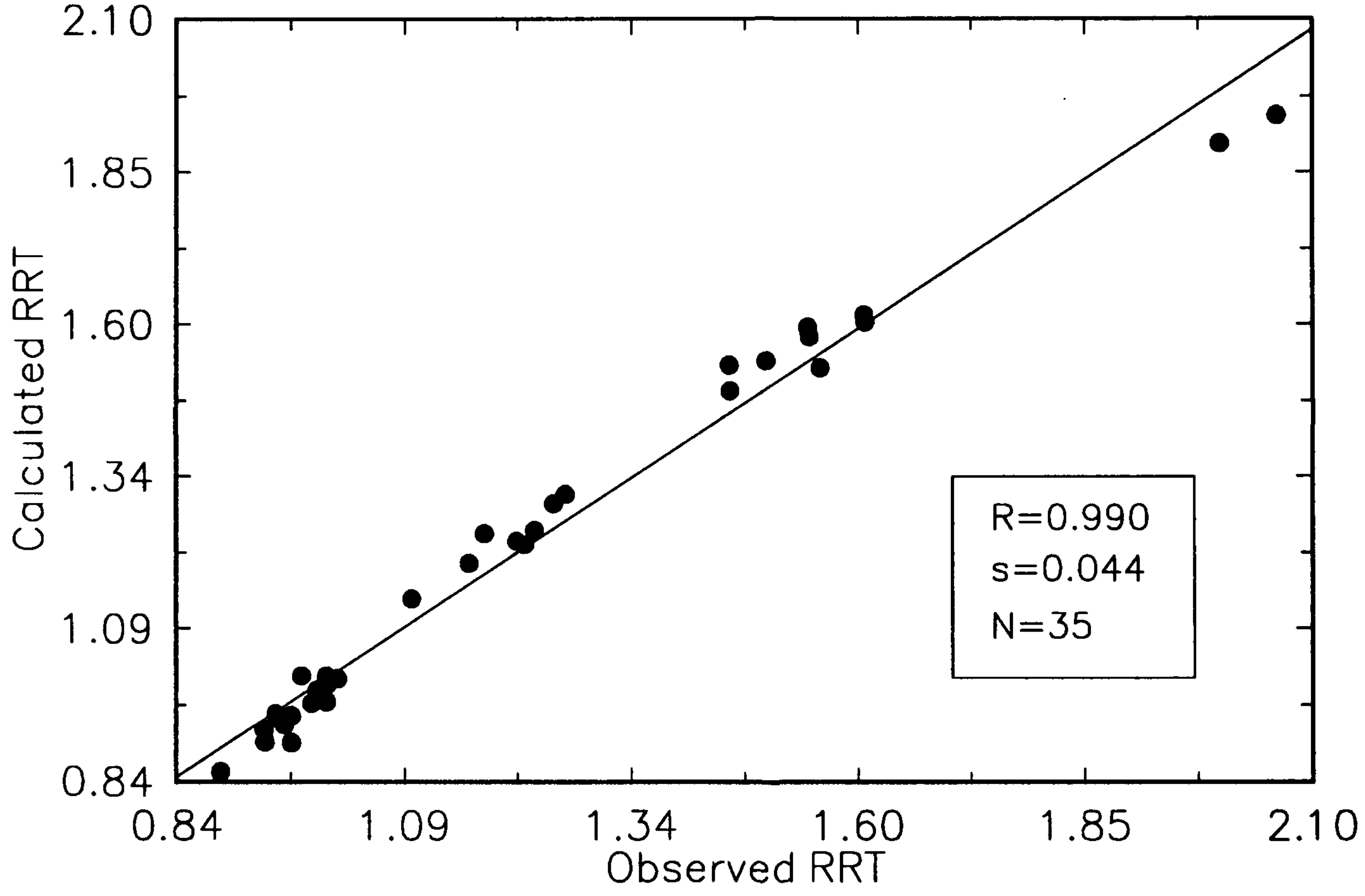


Figure 4.3 Calculated vs. observed plot for OV-101.

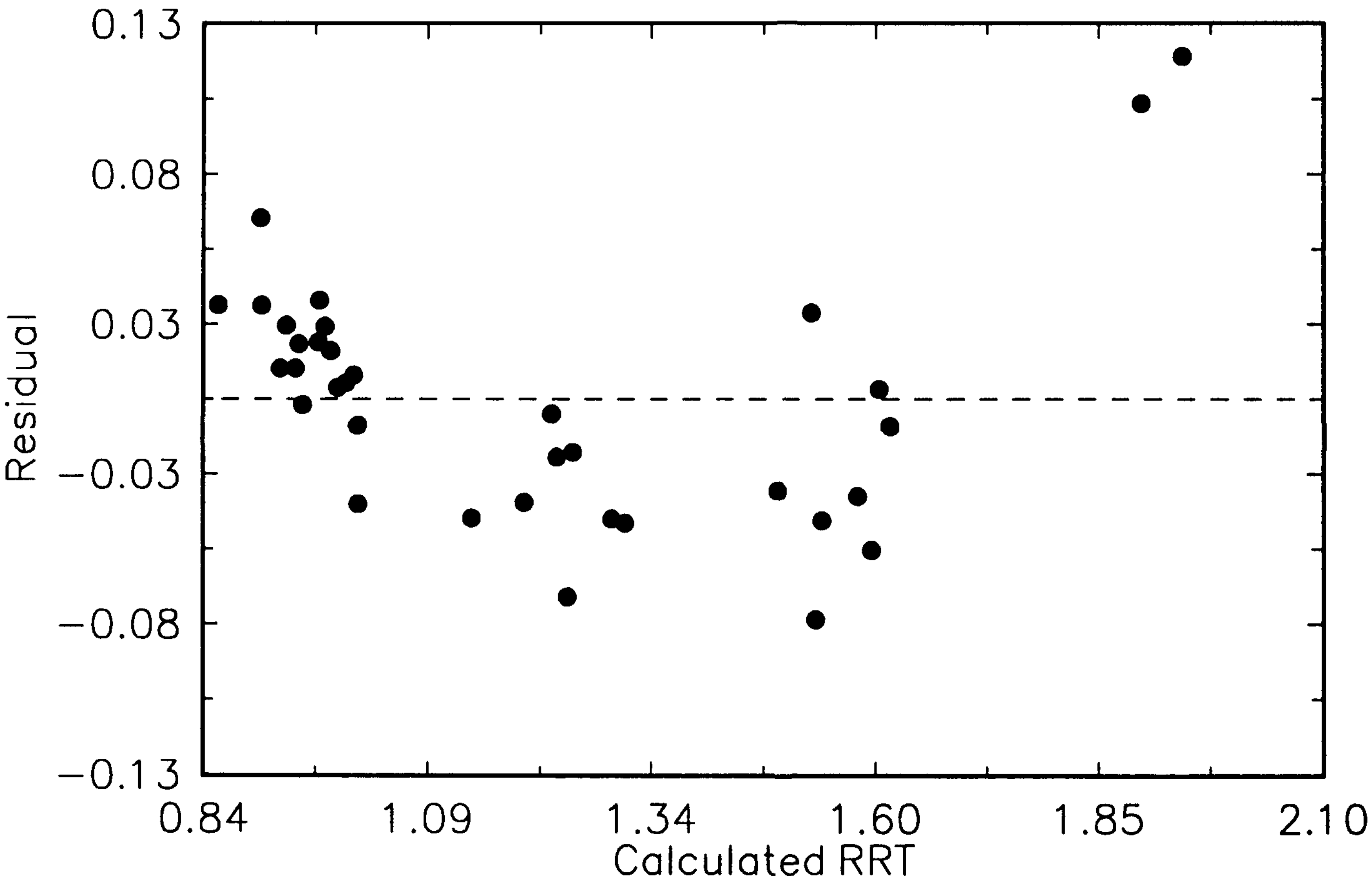


Figure 4.4 Residual plot for OV-101.

To start the process, a simple transformation of $Y_T=Y^{0.5}$ was attempted where Y_T is the transformed dependent variable. This square root transformation improved both the calculated versus observed plot (Figure 4.5) and the residual plot (Figure 4.6). Since this simple transformation was able to reduce the error to $s=0.013$ or about 1.0% at the mean of the range and improve the multiple correlation coefficient to $R=0.996$ it was decided that this model was sufficient to predict other relative retention times. As a further test, Box-Cox transformations (32) were attempted to determine the best transformation possible to minimize both the non-linearity and non-constant variance problems. The best transformation was calculated to be $Y_T=Y^{0.1}$. This truly minimized the sum-squared error, and the two plots of calculated vs. observed and residual vs. calculated were a significant improvement over the non-transformed plots. However, there was a problem with accepting this model as the best. Since the square root transformation already had an error of 1.0%, any model producing a standard error of less than this may be overfitting the data. The experimental error for this data set could not be assumed to be any better than about 1%; therefore, the square root model was chosen. During the modeling procedure, the independent variables were not transformed. New coefficients were generated and the model for OV-101 where $Y_T=Y^{0.5}$ is shown in equation 4. To make predictions, the model's calculated values were squared to transform them back to the proper units of minutes.

Outlier Analysis

Outliers are poorly fitted points which for some reason cannot be brought back into the fitted space without altering the model variables and/or compromising

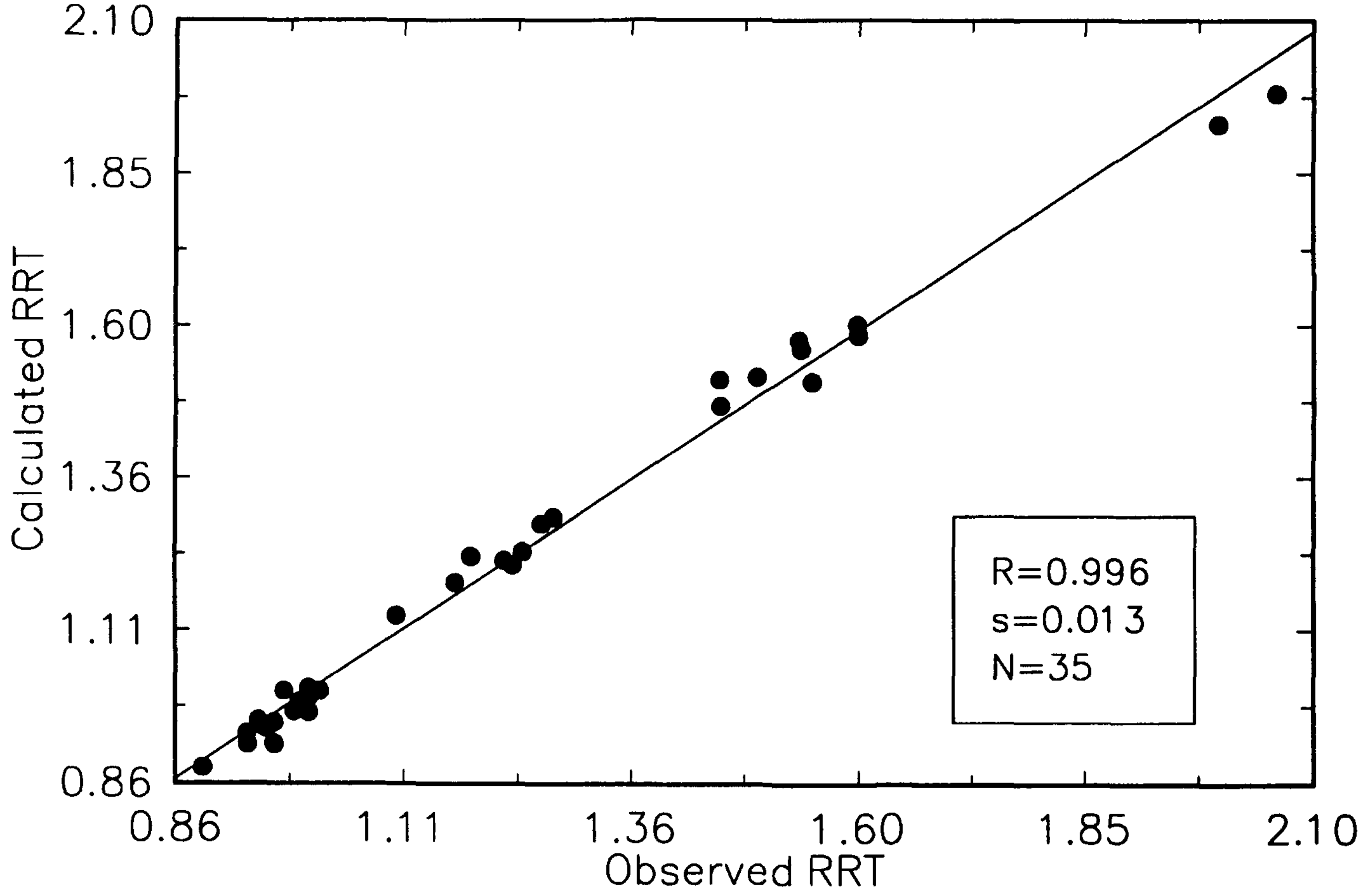


Figure 4.5 Calculated vs. observed plot for OV-101 after transformation.

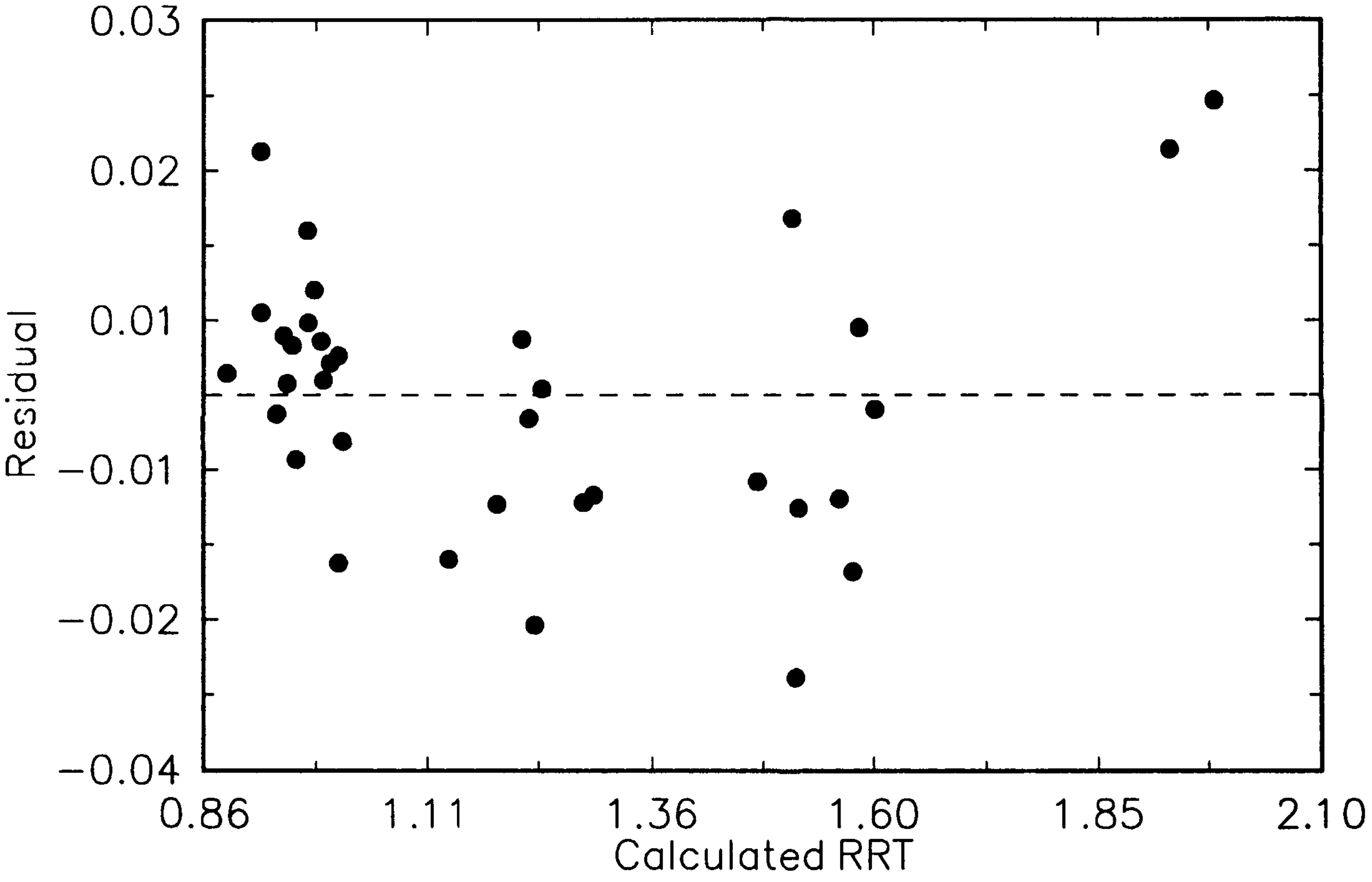


Figure 4.6 Residual plot for OV-101 after transformation.

$$\begin{aligned} RRT = {} & 0.007932 \quad \text{CSTR3 (2C)} \; + \\ & 0.001730 \quad \text{3D-W} \; + \\ & -0.3502 \quad \text{Const.} \end{aligned} \tag{4}$$

R=0 .996 $\quad s = 0.013(1.0\%)$ $\quad$ F= 1788 $\quad$ N=35

the validity of the model. Outliers can exist for many reasons, for instance, a poorly measured experimental value or an isomer which the descriptors cannot explain properly.

Two different techniques were used to check for outliers statistically as discussed in Chapter 2. The techniques were the data diagnostics generation (DDG) routine inherent to the ADAPT software system and robust regression analysis (RRA) which uses a least median squares instead of a least mean squares approach to regression analysis (33-35). Both methods were compared and outliers were identified on the basis of the results from both routines.

DDG uses five different tests--DFFITS, Cook's distance, leverage, studentized residuals and standardized residuals--to determine a point's fit. As described in Chapter 2 the routine will calculate a cutoff value for each of the tests and isomers exceeding the cutoff values for three of the five tests are considered to be an outlier. The outliers determined by DDG are shown in Table 4.5. The results show no significant problems with outliers. No one isomer was an outlier on more than one column. The data set with the most outliers was the DB-5 RI column which had three outliers. This is not significant considering the column had 115 observations at the start. The outliers were also spread throughout the range of isomers from mono-substituted to the octa-substituted which seems to verify that all

Table 4.5 Outliers determined by DDG.

Isomer	DB-5RI	DB-5RRT	OV-1701	OV-101	SE-54	SP-2330
1-monoCl		X				
18-diCl	X					
1234-teCl	X					
12367-peCl	X					
124689-heCl					X	
134678-heCl						X
234678-heCl						X
1234689-hpCl					X	
12346789-ocCl		X				

isomers are being predicted fairly well, i.e. no one group such as the tetrachloro isomers are falling out of the models.

The robust regression analysis method calculates a regression model using a least median squares algorithm which is not as susceptible to outlying data points. From the least median squares results, the program calculates the standard error, and if the error of any point exceeds 2.5 standard deviations, the point is considered to be an outlier. Outlying points then receive a weight of zero and a least mean squares is calculated without the outliers. RRA results are shown in Table 4.6. RRA found outliers on three of the five chromatographic columns. OV-1701 and OV-101 had points which were borderline outliers but not enough to be singled out as outliers. RRA did find a total of 16 points as outliers on the remaining four data sets which reflects the retention values for 13 isomers since three isomers were detected as outliers for two different columns. The isomers chosen twice were 1,2,3,6,8,9 hexachloro, 2,3,4,6,7,8 hexachloro and 1,2,3,4,6,8,9 heptachloro. The 1,2,3,6,8,9 hexachloro isomer was a borderline outlier for the SP-2330 and DB-5 RRT data sets. The models without these points were not very different, and the descriptor coefficients did not vary by more than one standard deviation. The 2,3,4,6,7,8 hexachloro isomer was an outlier for the SP-2330 and SE-54 data sets. It was a borderline outlier for the SE-54 data set but a true outlier for the SP-2330 column. This isomer did have the greatest retention time of any hexachloro isomer and a large leverage value in DDG test results. The 1,2,3,4,6,8,9 heptachloro isomer was also an outlier for the SP-2330 and SE-54 columns. As the point of greatest value of the experimentally available observations it did have a high leverage value for the SE-54 column. For the SP-2330 column its least median squares residual was greater than 2.5 standard deviations. The remaining isomers which RRA detected as

Table 4.6 Outliers determined by RRA.

Isomer	DB-5RI	DB-5RRT	OV-1701	OV-101	SE-54	SP-2330
1-monoCl		X				
18-diCl	X					
127-trCl	X					
1234-teCl	X					
1246-teCl	X					
12367-peCl	X					
123678-heCl		X				
123689-heCl		X				X
124678-heCl						X
134678-heCl						X
234678-heCl					X	X
1234678-hpCl					X	
1234689-hpCl					X	X

outliers were spread throughout the range of data from mono-substituted to the heptachloro isomers with no one particular substitution pattern standing out as a potential problem of the predictability of the models. These points could have been outliers for any of the reasons mentioned earlier.

When selecting isomers which would be excluded from modeling; both methods, DDG and RRA, were examined. The results are shown in Table 4.7. A total of 11 different isomers were identified as outliers and removed from consideration during model development. Only one isomer, 1,2,3,4,6,8,9 heptachloro, was an outlier on two different data sets. This was due to high leverage values. Seven data points were found to be outliers by both methods; three isomers were determined by RRA alone and one isomer was determined by DDG alone. No clear pattern was evident which would imply a problem with the models. All classes of isomers from mono-substituted to the octachloro isomer were being calculated very well. These points were true outliers and their exclusion improved the model statistics. Again two columns, OV-1701 and OV-101, had no outliers to exclude. The OV-101 column did have two data points for the two heaviest isomers which could have been selected as outliers. It was determined that the reason these points were borderline was because of the non-linearity problem discussed earlier. If these points had been deleted at the beginning of the modeling process, a non-linearity problem would have most likely never been detected. Since these points were included in calculating all transformations, they were left included in all models. Any exclusion of these points and transformations may not have been necessary.

The total number of outliers was minimal, but since these points were excluded from regression analysis new model coefficients had to be calculated. The results are shown in Table 4.8. The relative standard error shown as a percentage of

Table 4.7 Actual outliers removed and which outlier method(s) selected the isomer as an outlier.

Isomer	DB-5RI	DB-5RRT	OV-1701	OV-101	SE-54	SP-2330
1-monoCl		BOTH				
18-diCl	BOTH					
127-trCl	RRA					
1234-teCl	BOTH					
1246-teCl	RRA					
12367-peCl	BOTH					
134678-heCl						BOTH
234678-heCl						BOTH
1234678-hpCl					RRA	
1234689-hpCl					BOTH	RRA
12346789-ocCl		DDG				

Table 4.8 Summary of final models.

Descriptors	DB-5RI	DB-5RRT	OV-1701	OV-101	SE-54	SP-2330
MOLC1	534.0	0.457				
QSUM			1.652			
SHDW3						-0.0227
3D-W				0.00173	0.00140	0.00347
WPSA2	11.92			0.00793		
MOMH2	0.0553					
CSTR3(2C)			0.709		0.479	2.990
CSTR3(4C)						
TOAC1(2C)						6.573
TOHC3(2C)						-0.412
WHK2(2C)	-159.3					
MNAC1(4C)		-5.017				
CONST.	-1808	-2.405	-1.579	-0.350	0.281	-1.593
R	0.999	0.999	0.988	0.996	0.994	0.987
s	12.77(0.5%)	0.013(1.2%)	0.023(1.8%)	0.013(1.0%)	0.013(1.0%)	0.051(3.8%)
F	12496	4986	229	1788	388	194
N	110	24	14	35	12	32

the mean of the range is also presented. This shows that the error for most of the column is now very close to the assumed experimental error of 1.0%. Since the experimental error of the DB-5 RI column is known to be approximately seven retention index units or less than 0.3%, an error of 0.5% was considered excellent.

Model Validation

Validation of the regression models was achieved internally and primarily with the method known as jackknifing as explained in Chapter 2. However, for one of the data sets, DB-5 RI, another internal validation method called duplexing (36) was also used. Since this data set had 110 observations to use in the final modeling process and needed only four descriptors to describe the retention behavior, the data set was randomly divided in half numerous times and the model coefficients were recalculated using half the data. The duplexing models' coefficients were compared to the coefficients of the final model to determine if there were any large changes in the coefficients for the four descriptors. Any large changes (greater than one standard deviation) would imply a possible validity problem. The duplexing test; however, showed the DB-5 RI model to be extremely stable. Recalculated coefficients were very close to those of the final model and varied much less than one standard deviation.

There were not as many experimental observations available for the other data sets and jackknifing was used as the internal test of the models validity. Jackknifing recalculates the model numerous times with each observation held out once. It then uses this new model to generate a predicted value for all of the points when that point is not included in the modeling process. The jackknifing results showed no validation problems for any of the models. All jackknifed estimates were extremely close to the calculated values using all available observations in the modeling process.

In a final analysis, plots of calculated versus observed and residuals were generated. Figures 4.7 through 4.12 show the calculated versus observed plots for all six data sets. These plots further demonstrate the high quality of the models. The plots show an excellent fit of the experimental data to the calculated values.

Predictions

Since it was not possible to generate models using experimental observations representing all classes of isomers for each data set, it would not be valid to make predictions in these areas. For instance, there were no mono-, di- or tri-substituted experimental observations available for the OV-101, OV-1701, SE-54 and SP-2330 data sets, therefore, no predictions were made for these classes of compounds for the above stationary phases. However, predictions of the remaining isomers were generated. This rule was enforced in order to avoid overstepping the bounds of the regression models. Predicted values were determined for all isomers of the two DB-5 data sets. Observed, predicted and residual values for all data sets are shown in Table 4.9.

Predictions were also determined for the outliers detected previously. This was done to provide a value the model would calculate for that point. A comparison of nearby isomers can be made and the predicted value need not be accepted.

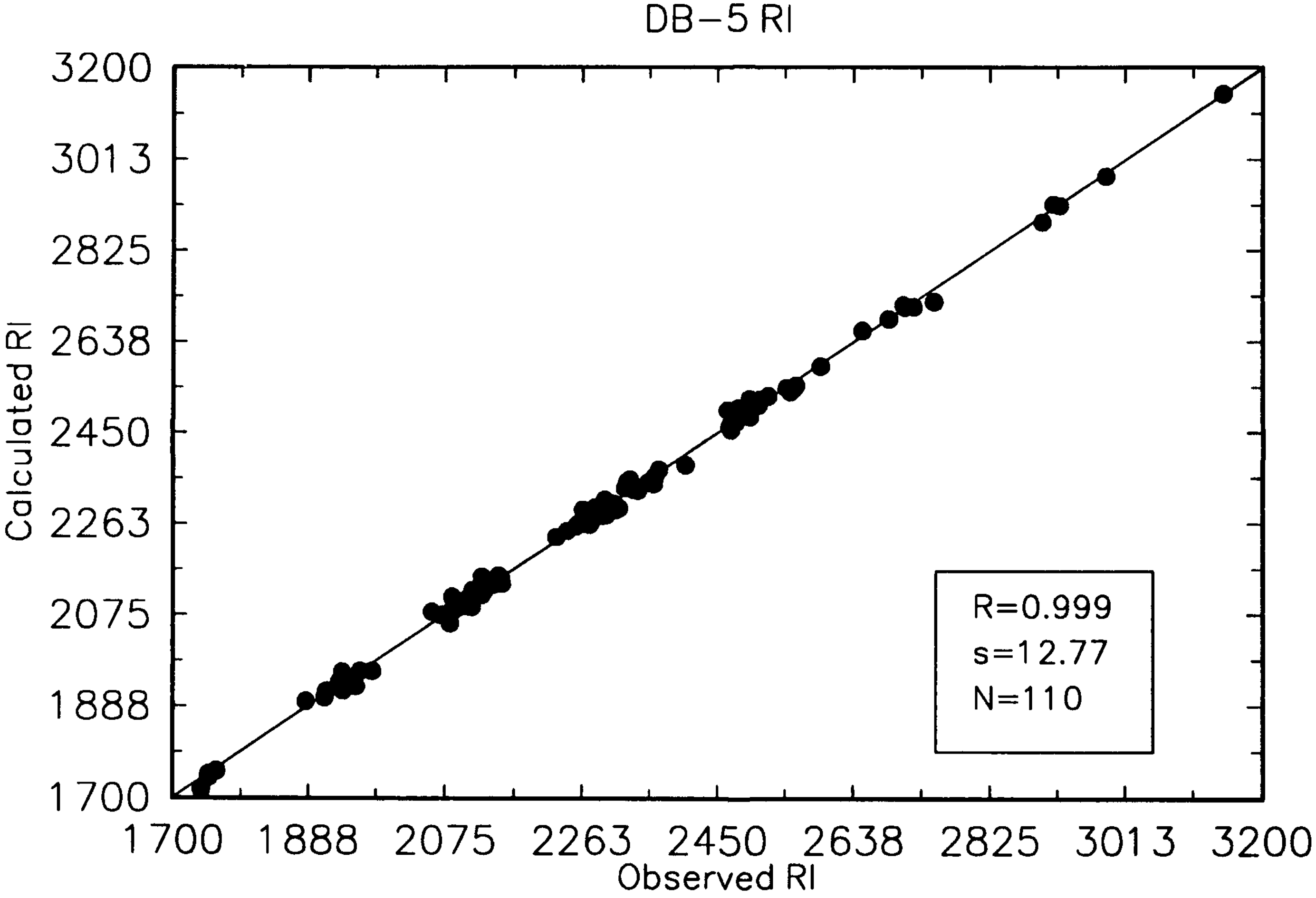


Figure 4.7 Calculated vs. observed plot for the DB-5 RI data set.

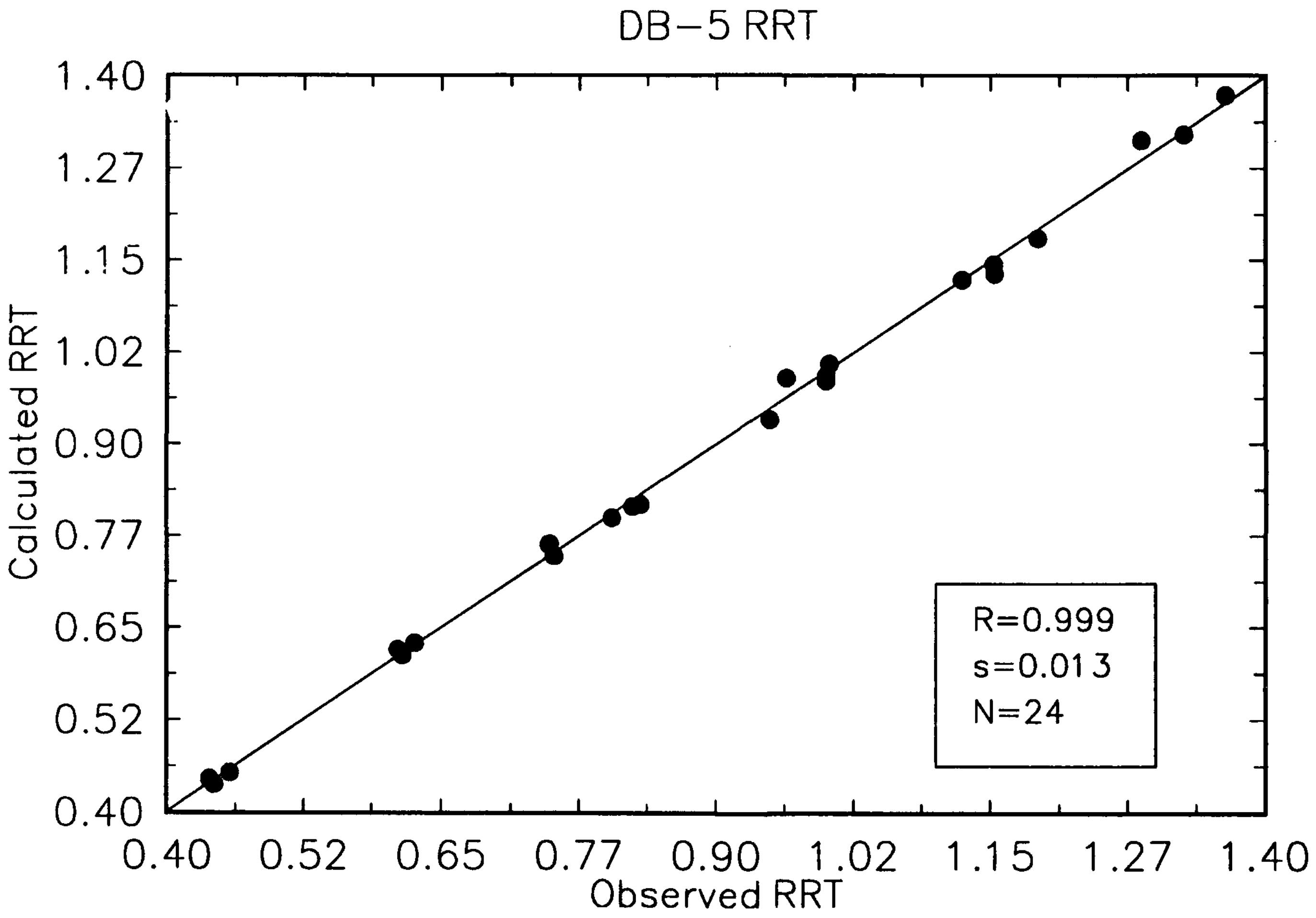


Figure 4.8 Calculated vs. observed plot for the DB-5 RRT data set.

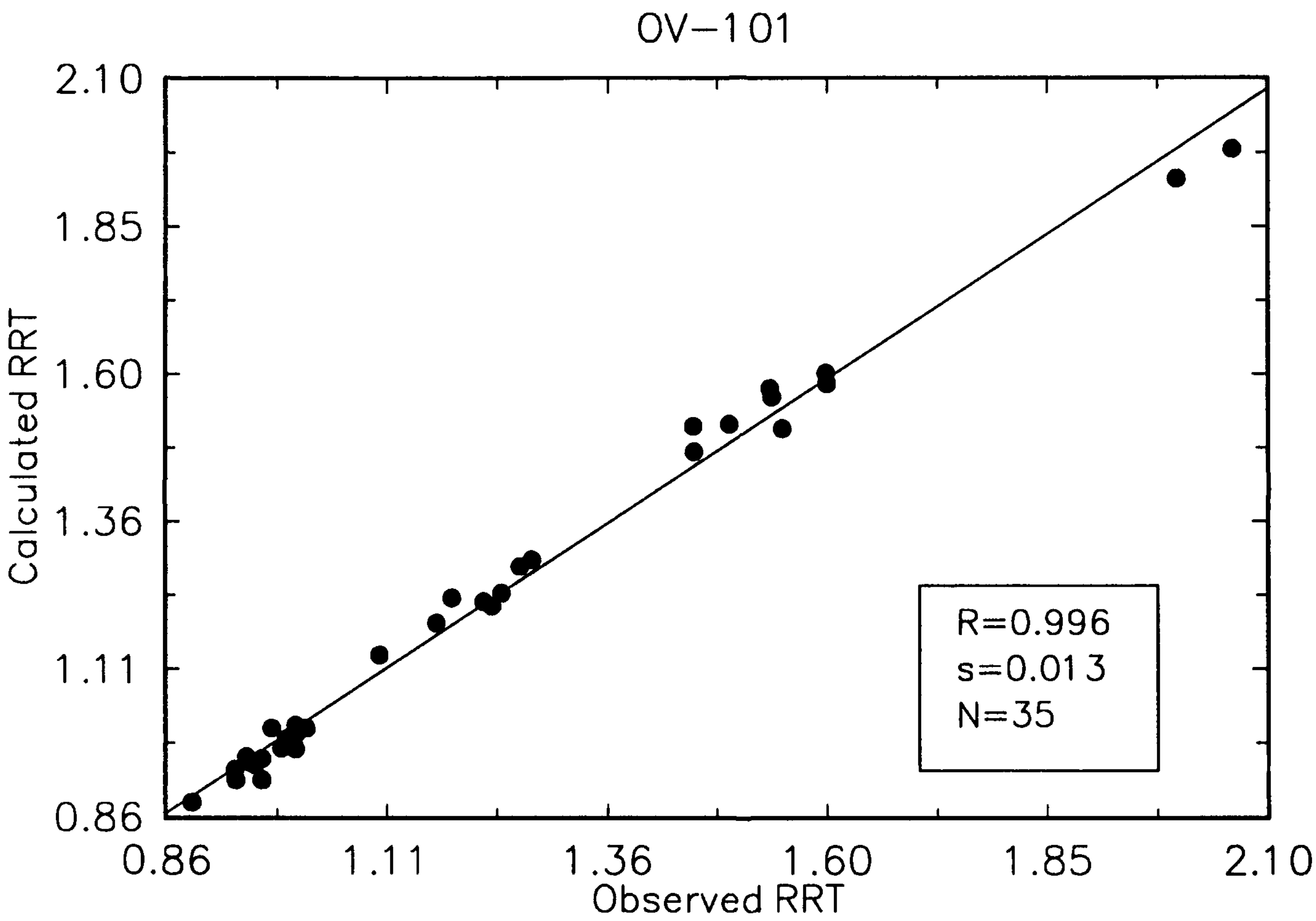


Figure 4.9 Calculated vs. observed plot for the OV-101 data set.

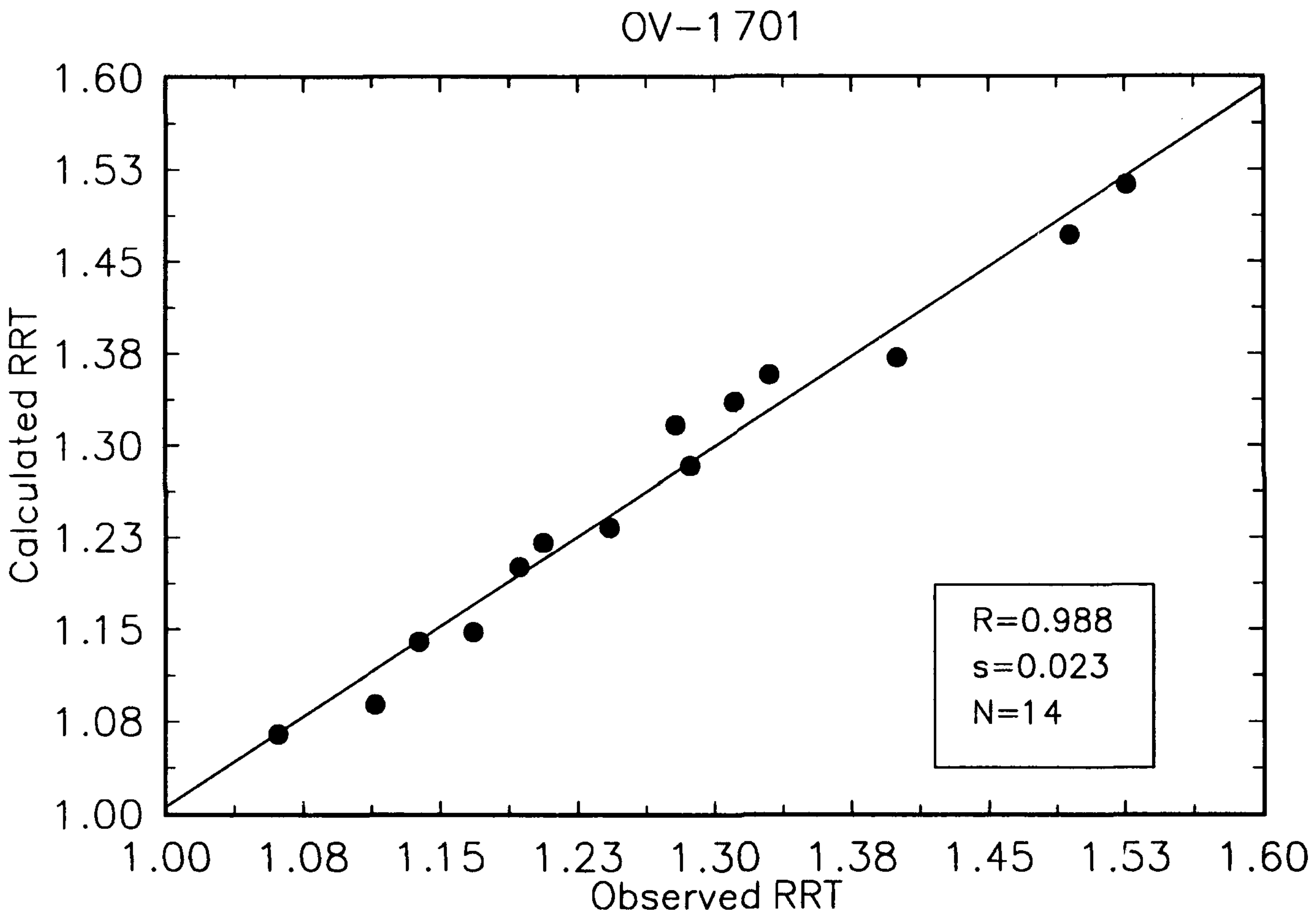


Figure 4.10 Calculated vs. observed plot for the OV-1701 data set.

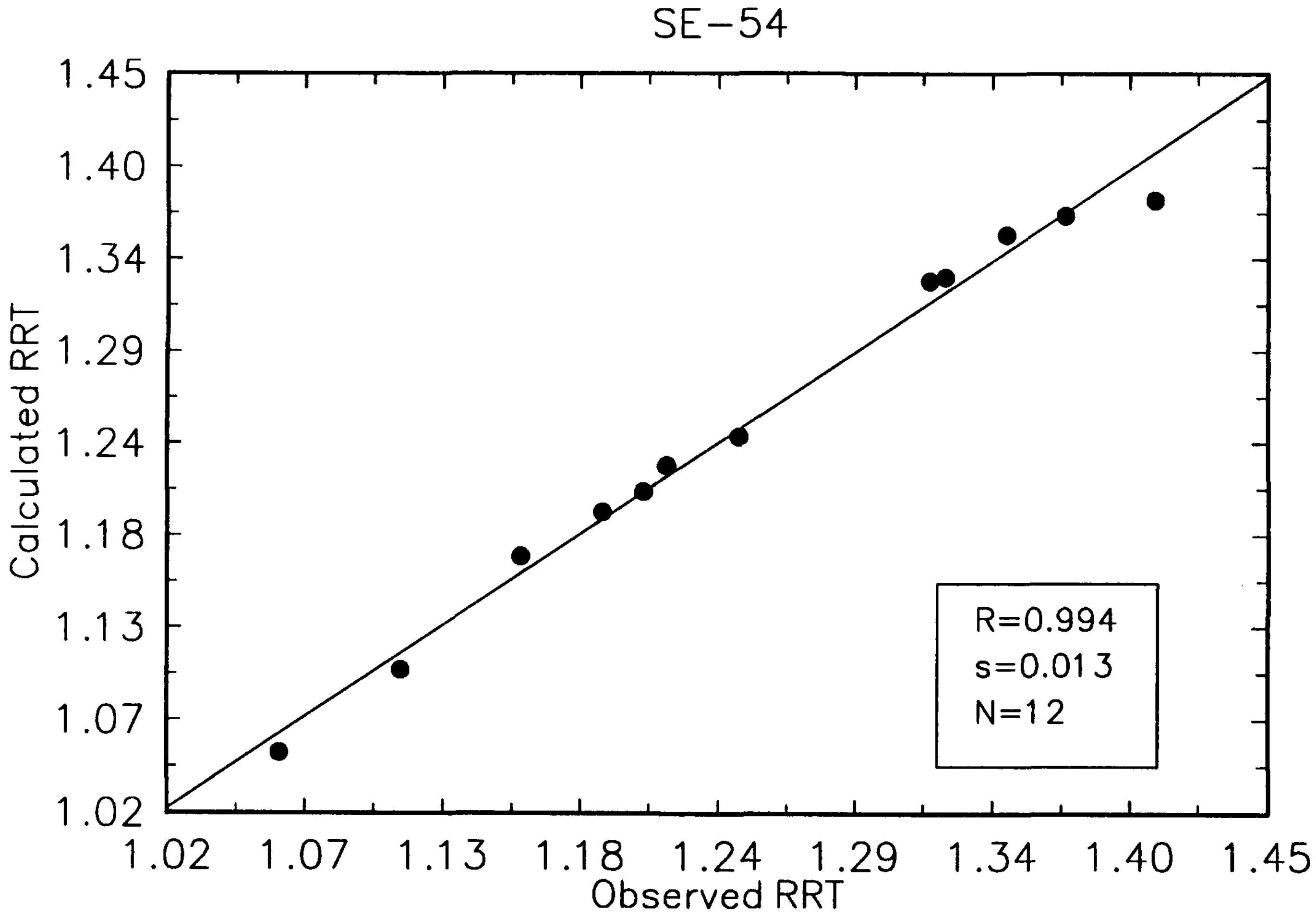


Figure 4.11 Calculated vs. observed plot for the SE-54 data set.

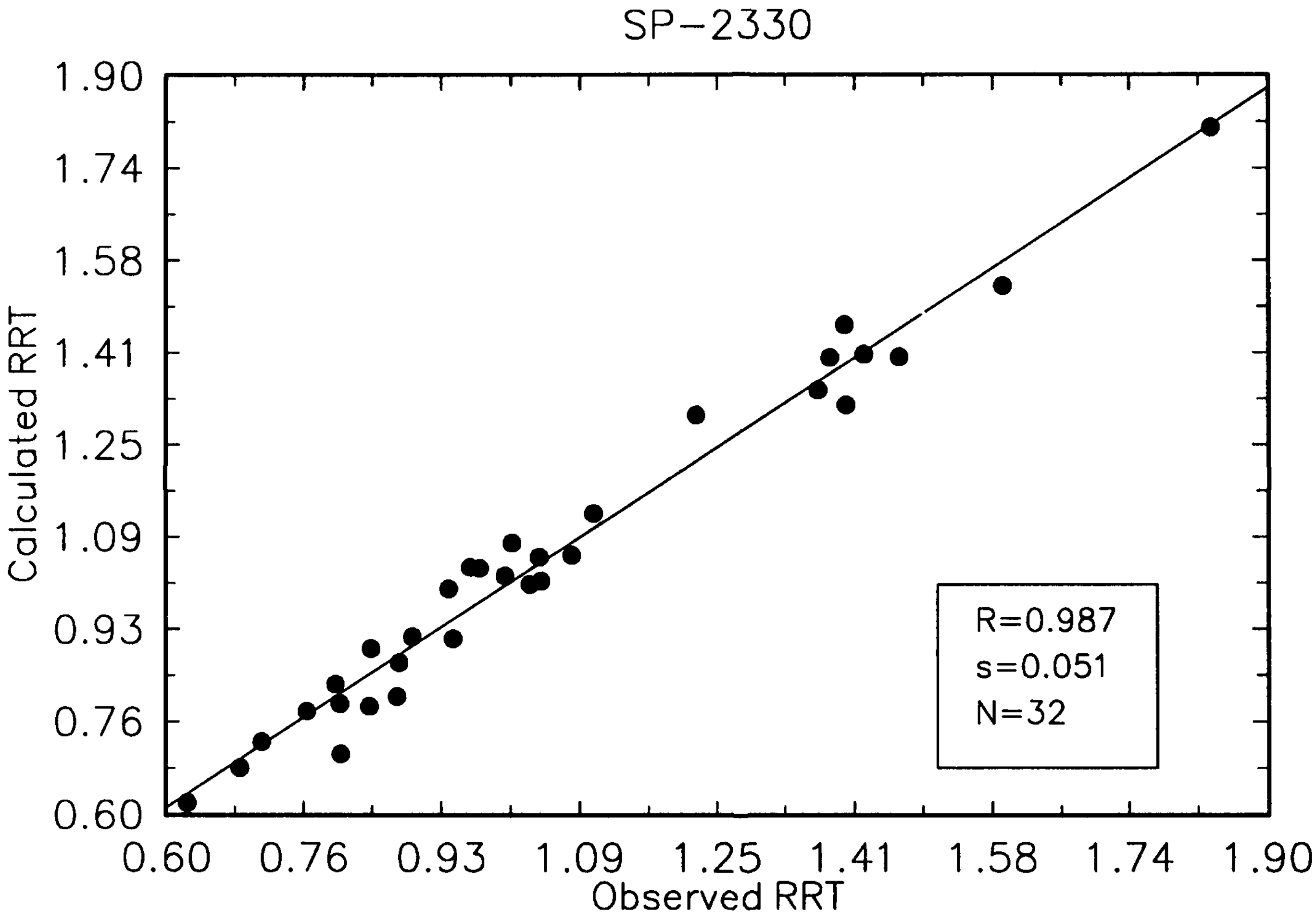


Figure 4.12 Calculated vs. observed plot for the SP-2330 data set.

Table 4.9 Predicted values for all data sets.

	DB-5 RI		
Isomer	Predicted	Observed	Residual
1-Cl	1717	1739	22
2-Cl	1742	1749	7
3-Cl	1749	1749	0
4-Cl	1755	1760	5
12-diCl	1958	1934	-24
13-diCl	1898	1884	-14
14-diCl	1919	1913	-6
16-diCl	1915		
17-diCl	1905	1910	5
18-diCl	1891	1925	34
19-diCl	1960	1975	15
23-diCl	1945	1939	-6
24-diCl	1918	1912	-6
26-diCl	1945	1946	1
27-diCl	1936	1930	-6
28-diCl	1919	1935	16
34-diCl	1960	1959	-1
36-diCl	1938	1944	6
37-diCl	1938	1930	-8
46-diCl	1928	1953	25
123-trCl	2127	2113	-14
124-trCl	2112	2085	-27
126-trCl	2153	2125	-28
127-trCl	2144	2109	-35
128-trCl	2126	2129	3
129-trCl	2171		
134-trCl	2093	2088	-5
136-trCl	2076	2072	-4
137-trCl	2082	2057	-25
138-trCl	2074	2070	-4
139-trCl	2122	2124	2
146-trCl	2091	2094	3
147-trCl	2083	2086	3
148-trCl	2094	2100	6

Table 4.9 (Cont.)

	DB-5 RI		
Isomer	Predicted	Observed	Residual
149-trCl	2148	2151	3
234-trCl	2156	2148	-8
236-trCl	2138	2141	3
237-trCl	2139	2134	-5
238-trCl	2132	2132	0
239-trCl	2091	2111	20
246-trCl	2105	2101	-4
247-trCl	2103	2099	-4
248-trCl	2094	2097	3
249-trCl	2058	2082	24
346-trCl	2139	2152	13
347-trCl	2146	2150	4
348-trCl	2151	2151	0
349-trCl	2116	2125	9
1234-teCl	2345	2310	-35
1236-teCl	2303	2307	4
1237-teCl	2312	2294	-18
1238-teCl	2295	2307	12
1239-teCl	2374	2369	-5
1246-teCl	2296	2264	-32
1247-teCl	2292	2264	-28
1248-teCl	2283	2274	-9
1249-teCl	2332	2335	3
1267-teCl	2354	2329	-25
1268-teCl	2297	2281	-16
1269-teCl	2361	2364	3
1278-teCl	2336	2322	-14
1279-teCl	2335	2341	6
1289-teCl	2384	2406	22
1346-teCl	2263	2262	-1
1347-teCl	2263	2257	-6
1348-teCl	2269	2276	7
1349-teCl	2350	2325	-25
1367-teCl	2279	2272	-7

Table 4.9 (Cont.)

	DB-5 RI		
Isomer	Predicted	Observed	Residual
1368-teCl	2236	2227	-9
1369-teCl	2281	2296	15
1378-teCl	2267	2263	-4
1379-teCl	2261	2273	12
1467-teCl	2288	2288	0
1468-teCl	2248	2242	-6
1469-teCl	2295	2314	19
1478-teCl	2279	2290	11
2346-teCl	2334	2339	5
2347-teCl	2339	2337	-2
2348-teCl	2331	2340	9
2349-teCl	2291	2308	17
2367-teCl	2348	2354	6
2368-teCl	2295	2297	2
2378-teCl	2337	2338	1
2467-teCl	2306	2305	-1
2468-teCl	2257	2254	-3
3467-teCl	2345	2362	17
12346-peCl	2515	2496	-19
12347-peCl	2520	2495	-25
12348-peCl	2518	2508	-10
12349-peCl	2578		
12367-peCl	2505	2540	35
12368-peCl	2449		
12369-peCl	2543	2546	3
12378-peCl	2508	2507	-1
12379-peCl	2525		
12389-peCl	2588	2593	5
12467-peCl	2497	2465	-32
12468-peCl	2436		
12469-peCl	2498	2497	-1
12478-peCl	2478		
12479-peCl	2501	2479	-22
12489-peCl	2548	2559	11

Table 4.9 (Cont.)

Isomer	DB-5 RI Predicted	Observed	Residual
13467-peCl	2457	2469	12
13468-peCl	2414		
13469-peCl	2495		
13478-peCl	2458	2469	11
13479-peCl	2477	2473	-4
13489-peCl	2565		
23467-peCl	2541	2555	14
23468-peCl	2483	2495	12
23469-peCl	2472	2476	4
23478-peCl	2535	2551	16
23479-peCl	2463	2467	4
23489-peCl	2527	2521	-6
123467-heCl	2714	2706	-8
123468-heCl	2661	2650	-11
123469-heCl	2723		
123478-heCl	2709	2708	-1
123479-heCl	2710	2720	10
123489-heCl	2792		
123678-heCl	2681		
123679-heCl	2743		
123689-heCl	2733		
123789-heCl	2799		
124678-heCl	2670		
124679-heCl	2695		
124689-heCl	2685	2686	1
134678-heCl	2635		
134679-heCl	2688		
234678-heCl	2720	2748	28
1234678-hpCl	2884	2898	14
1234679-hpCl	2921	2913	-8
1234689-hpCl	2918	2922	4
1234789-hpCl	2978	2986	8
12346789-ocCl	3147	3147	0

Table 4.9 (Cont.)

	DB-5 RRT		
Isomer	Predicted	Observed	Residual
1-Cl	0.387	0.341	-0.046
2-Cl	0.438	0.443	0.005
3-Cl	0.446	0.439	-0.007
4-Cl	0.454	0.457	0.003
12-diCl	0.579		
13-diCl	0.577		
14-diCl	0.571		
16-diCl	0.575		
17-diCl	0.574		
18-diCl	0.558		
19-diCl	0.472		
23-diCl	0.625		
24-diCl	0.629		
26-diCl	0.629	0.626	-0.003
27-diCl	0.621	0.611	-0.010
28-diCl	0.613	0.615	0.002
34-diCl	0.640		
36-diCl	0.637		
37-diCl	0.629		
46-diCl	0.646		
123-trCl	0.771		
124-trCl	0.763		
126-trCl	0.766		
127-trCl	0.766		
128-trCl	0.750		
129-trCl	0.660		
134-trCl	0.764		
136-trCl	0.764	0.748	-0.016
137-trCl	0.763	0.747	-0.016
138-trCl	0.748	0.752	0.004
139-trCl	0.665		
146-trCl	0.759		
147-trCl	0.758		
148-trCl	0.742		

Table 4.9 (Cont.)

Isomer	DB-5 RRT Predicted	Observed	Residual
149-trCl	0.652		
234-trCl	0.818	0.831	0.013
236-trCl	0.816		
237-trCl	0.808		
238-trCl	0.800	0.805	0.005
239-trCl	0.748		
246-trCl	0.821		
247-trCl	0.812		
248-trCl	0.804		
249-trCl	0.745		
346-trCl	0.832		
347-trCl	0.823		
348-trCl	0.815	0.824	0.009
349-trCl	0.764		
1234-teCl	0.958		
1236-teCl	0.959		
1237-teCl	0.958		
1238-teCl	0.942		
1239-teCl	0.855		
1246-teCl	0.951		
1247-teCl	0.950		
1248-teCl	0.934	0.949	0.015
1249-teCl	0.840		
1267-teCl	0.956		
1268-teCl	0.937		
1269-teCl	0.840		
1278-teCl	0.940		
1279-teCl	0.853		
1289-teCl	0.848		
1346-teCl	0.951		
1347-teCl	0.950		
1348-teCl	0.935		
1349-teCl	0.847		
1367-teCl	0.953		

Table 4.9 (Cont.)

Isomer	DB-5 RRT Predicted	Observed	Residual
1368-teCl	0.934		
1369-teCl	0.845		
1378-teCl	0.938		
1379-teCl	0.857		
1467-teCl	0.948		
1468-teCl	0.929		
1469-teCl	0.832		
1478-teCl	0.932		
2346-teCl	1.010	1.003	-0.007
2347-teCl	1.001		
2348-teCl	0.994	1.000	0.006
2349-teCl	0.938		
2367-teCl	1.002		
2368-teCl	0.991	0.964	-0.027
2378-teCl	0.987	1.000	0.013
2467-teCl	1.006		
2468-teCl	0.995		
3467-teCl	1.017		
12346-peCl	1.146		
12347-peCl	1.145	1.153	0.008
12348-peCl	1.129		
12349-peCl	1.038		
12367-peCl	1.148		
12368-peCl	1.129		
12369-peCl	1.035		
12378-peCl	1.132	1.154	0.022
12379-peCl	1.048		
12389-peCl	1.043		
12467-peCl	1.140		
12468-peCl	1.121		
12469-peCl	1.020		
12478-peCl	1.124	1.124	0.000
12479-peCl	1.032		
12489-peCl	1.028		

Table 4.9 (Cont.)

Isomer	DB-5 RRT Predicted	Observed	Residual
13467-peCl	1.140		
13468-peCl	1.121		
13469-peCl	1.027		
13478-peCl	1.125		
13479-peCl	1.040		
13489-peCl	1.035		
23467-peCl	1.196		
23468-peCl	1.185		
23469-peCl	1.122		
23478-peCl	1.180	1.193	0.013
23479-peCl	1.127		
23489-peCl	1.129		
123467-heCl	1.335		
123468-heCl	1.316		
123469-heCl	1.218		
123478-heCl	1.319		
123479-heCl	1.231		
123489-heCl	1.226		
123678-heCl	1.321	1.326	0.005
123679-heCl	1.230		
123689-heCl	1.223		
123789-heCl	1.239		
124678-heCl	1.313	1.287	-0.026
124679-heCl	1.215		
124689-heCl	1.208		
134678-heCl	1.314		
134679-heCl	1.223		
234678-heCl	1.374	1.364	-0.010
1234678-hpCl	1.508		
1234679-hpCl	1.413		
1234689-hpCl	1.406		
1234789-hpCl	1.421		
12346789-ocCl	1.604	1.798	0.194

Table 4.9 (Cont.)

Isomer	SP-2330 Predicted	Observed	Residual
1234-teCl	0.828	0.800	-0.028
1236-teCl	0.671		
1237-teCl	0.782	0.766	-0.016
1238-teCl	0.794	0.805	0.011
1239-teCl	0.938		
1246-teCl	0.622		
1247-teCl	0.650		
1248-teCl	0.662		
1249-teCl	0.827		
1267-teCl	0.807	0.873	0.066
1268-teCl	0.690		
1269-teCl	0.868		
1278-teCl	0.790	0.840	0.050
1279-teCl	0.867	0.875	0.008
1289-teCl	0.992		
1346-teCl	0.652		
1347-teCl	0.686		
1348-teCl	0.683		
1349-teCl	0.813		
1367-teCl	0.727	0.713	-0.014
1368-teCl	0.621	0.625	0.004
1369-teCl	0.608		
1378-teCl	0.729		
1379-teCl	0.682	0.687	0.005
1467-teCl	0.707	0.806	0.099
1468-teCl	0.588		
1469-teCl	0.695		
1478-teCl	0.596		
2346-teCl	1.004	1.029	0.026

Table 4.9 (Cont.)

	SP-2330		
Isomer	Predicted	Observed	Residual
2347-teCl	1.033	0.970	-0.063
2348-teCl	1.077		
2349-teCl	0.782		
2367-teCl	1.010	1.042	0.032
2368-teCl	0.912	0.891	-0.021
2378-teCl	1.019	1.000	-0.019
2467-teCl	0.996	0.934	-0.062
2468-teCl	0.883		
3467-teCl	1.128		
12346-peCl	1.020		
12347-peCl	1.052		
12348-peCl	1.083		
12349-peCl	1.210		
12367-peCl	1.056	1.078	0.022
12368-peCl	0.938		
12369-peCl	1.085		
12378-peCl	1.052	1.040	-0.012
12379-peCl	1.113		
12389-peCl	1.339		
12467-peCl	0.987		
12468-peCl	0.892	0.842	-0.050
12469-peCl	1.053		
12478-peCl	0.908	0.939	0.031
12479-peCl	1.034	0.959	-0.075
12489-peCl	1.228		
13467-peCl	1.003		
13468-peCl	0.932		
13469-peCl	0.980		
13478-peCl	0.963		

Table 4.9 (Cont.)

Isomer	SP-2330 Predicted	Observed	Residual
13479-peCl	0.947		
13489-peCl	1.164		
23467-peCl	1.406	1.465	0.059
23468-peCl	1.287		
23469-peCl	0.962		
23478-peCl	1.321	1.403	0.082
23479-peCl	1.026		
23489-peCl	1.129	1.104	-0.025
123467-heCl	1.410	1.424	0.014
123468-heCl	1.302		
123469-heCl	1.491		
123478-heCl	1.347	1.370	0.023
123479-heCl	1.485		
123489-heCl	1.573		
123678-heCl	1.404	1.384	-0.020
123679-heCl	1.485		
123689-heCl	1.531	1.587	0.056
123789-heCl	1.611		
124678-heCl	1.303	1.225	-0.078
124679-heCl	1.272		
124689-heCl	1.462	1.401	-0.061
134678-heCl	1.342	1.199	-0.143
134679-heCl	1.334		
234678-heCl	1.729	2.001	0.272
1234678-hpCl	1.809	1.834	0.025
1234679-hpCl	1.902		
1234689-hpCl	1.925	2.084	0.159
1234789-hpCl	2.002		

Table 4.9 (Cont.)

	SE-54		
Isomer	Predicted	Observed	Residual
1234-teCl	1.065		
1236-teCl	1.060		
1237-teCl	1.082		
1238-teCl	1.073		
1239-teCl	1.095		
1246-teCl	1.032		
1247-teCl	1.058		
1248-teCl	1.050		
1249-teCl	1.072		
1267-teCl	1.071		
1268-teCl	1.053		
1269-teCl	1.085		
1278-teCl	1.075		
1279-teCl	1.089		
1289-teCl	1.090		
1346-teCl	1.039		
1347-teCl	1.065		
1348-teCl	1.056		
1349-teCl	1.076		
1367-teCl	1.068		
1368-teCl	1.053		
1369-teCl	1.062		
1378-teCl	1.074		
1379-teCl	1.077		
1467-teCl	1.045		
1468-teCl	1.031		
1469-teCl	1.053		
1478-teCl	1.054		
2346-teCl	1.066		

Table 4.9 (Cont.)

Isomer	SE-54 Predicted	Observed	Residual
2347-teCl	1.094		
2348-teCl	1.091		
2349-teCl	1.058		
2367-teCl	1.092		
2368-teCl	1.083		
2378-teCl	1.103	1.111	0.008
2467-teCl	1.065		
2468-teCl	1.055	1.064	0.009
3467-teCl	1.077		
12346-peCl	1.183		
12347-peCl	1.211		
12348-peCl	1.206		
12349-peCl	1.224		
12367-peCl	1.214		
12368-peCl	1.196		
12369-peCl	1.227		
12378-peCl	1.222	1.215	-0.007
12379-peCl	1.237		
12389-peCl	1.252		
12467-peCl	1.183		
12468-peCl	1.169	1.158	-0.011
12469-peCl	1.198		
12478-peCl	1.195	1.190	-0.005
12479-peCl	1.215		
12489-peCl	1.230		
13467-peCl	1.186		
13468-peCl	1.177		
13469-peCl	1.192		
13478-peCl	1.207		

Table 4.9 (Cont.)

	SE-54		
Isomer	Predicted	Observed	Residual
13479-peCl	1.209		
13489-peCl	1.226		
23467-peCl	1.224		
23468-peCl	1.207	1.206	-0.001
23469-peCl	1.177		
23478-peCl	1.239	1.243	0.004
23479-peCl	1.206		
23489-peCl	1.210		
123467-heCl	1.346		
123468-heCl	1.330	1.318	-0.012
123469-heCl	1.363		
123478-heCl	1.362		
123479-heCl	1.387		
123489-heCl	1.382		
123678-heCl	1.368	1.371	0.003
123679-heCl	1.391		
123689-heCl	1.392		
123789-heCl	1.413		
124678-heCl	1.332	1.324	-0.008
124679-heCl	1.334		
124689-heCl	1.357	1.348	-0.009
134678-heCl	1.339		
134679-heCl	1.349		
234678-heCl	1.377	1.406	0.029
1234678-hpCl	1.514	1.567	0.053
1234679-hpCl	1.538		
1234689-hpCl	1.535	1.598	0.063
1234789-hpCl	1.573		

Table 4.9 (Cont.)

	OV-1701		
Isomer	Predicted	Observed	Residual
1234-teCl	0.315		
1236-teCl	0.235		
1237-teCl	0.192		
1238-teCl	0.135		
1239-teCl	0.301		
1246-teCl	0.170		
1247-teCl	0.138		
1248-teCl	0.074		
1249-teCl	0.253		
1267-teCl	0.209		
1268-teCl	0.098		
1269-teCl	0.242		
1278-teCl	0.102		
1279-teCl	0.217		
1289-teCl	0.220		
1346-teCl	0.230		
1347-teCl	0.194		
1348-teCl	0.133		
1349-teCl	0.292		
1367-teCl	0.228		
1368-teCl	0.123		
1369-teCl	0.208		
1378-teCl	0.126		
1379-teCl	0.196		
1467-teCl	0.215		
1468-teCl	0.103		
1469-teCl	0.227		
1478-teCl	0.110		
2346-teCl	0.250		

Table 4.9 (Cont.)

Isomer	OV-1701 Predicted	Observed	Residual
2347-teCl	0.221		
2348-teCl	0.163		
2349-teCl	0.235		
2367-teCl	0.189		
2368-teCl	0.082		
2378-teCl	0.089	0.114	0.025
2467-teCl	0.181		
2468-teCl	0.065	0.061	-0.004
3467-teCl	0.289		
12346-peCl	0.352		
12347-peCl	0.314		
12348-peCl	0.261		
12349-peCl	0.425		
12367-peCl	0.310		
12368-peCl	0.205		
12369-peCl	0.334		
12378-peCl	0.201	0.193	-0.008
12379-peCl	0.309		
12389-peCl	0.342		
12467-peCl	0.244		
12468-peCl	0.140	0.138	-0.002
12469-peCl	0.281		
12478-peCl	0.148	0.168	0.020
12479-peCl	0.276		
12489-peCl	0.294		
13467-peCl	0.297		
13468-peCl	0.204		
13469-peCl	0.306		
13478-peCl	0.206		

Table 4.9 (Cont.)

Isomer	OV-1701 Predicted	Observed	Residual
13479-peCl	0.293		
13489-peCl	0.323		
23467-peCl	0.334		
23468-peCl	0.221	0.206	-0.015
23469-peCl	0.250		
23478-peCl	0.233	0.242	0.009
23479-peCl	0.268		
23489-peCl	0.259		
123467-heCl	0.420		
123468-heCl	0.317	0.279	-0.038
123469-heCl	0.459		
123478-heCl	0.322		
123479-heCl	0.448		
123489-heCl	0.448		
123678-heCl	0.358	0.330	-0.028
123679-heCl	0.424		
123689-heCl	0.397		
123789-heCl	0.426		
124678-heCl	0.284	0.287	0.003
124679-heCl	0.337		
124689-heCl	0.336	0.311	-0.025
134678-heCl	0.343		
134679-heCl	0.389		
234678-heCl	0.372	0.400	0.028
1234678-hpCl	0.472	0.495	0.023
1234679-hpCl	0.548		
1234689-hpCl	0.513	0.526	0.013
1234789-hpCl	0.566		

Table 4.9 (Cont.)

Isomer	OV-101 Predicted	Observed	Residual
1234-teCl	1.009	0.978	-0.031
1236-teCl	0.939		
1237-teCl	0.962	0.950	-0.012
1238-teCl	0.923	0.967	0.044
1239-teCl	1.035		
1246-teCl	0.932		
1247-teCl	0.941		
1248-teCl	0.918		
1249-teCl	0.984		
1267-teCl	0.992	0.995	0.003
1268-teCl	0.924		
1269-teCl	0.995		
1278-teCl	0.976	0.989	0.013
1279-teCl	0.974	1.005	0.031
1289-teCl	1.014		
1346-teCl	0.922		
1347-teCl	0.925		
1348-teCl	0.902		
1349-teCl	1.018		
1367-teCl	0.940	0.937	-0.003
1368-teCl	0.885	0.889	0.004
1369-teCl	0.950		
1378-teCl	0.910		
1379-teCl	0.923	0.938	0.015
1467-teCl	0.952	0.954	0.002
1468-teCl	0.896		
1469-teCl	0.966		
1478-teCl	0.929		
2346-teCl	1.000	1.006	0.006

Table 4.9 (Cont.)

Isomer	OV-101 Predicted	Observed	Residual
2347-teCl	1.014	1.005	-0.009
2348-teCl	0.982	1.002	0.020
2349-teCl	0.942		
2367-teCl	1.009	1.017	0.008
2368-teCl	0.948	0.959	0.011
2378-teCl	0.990	1.000	0.010
2467-teCl	0.958	0.967	0.009
2468-teCl	0.903		
3467-teCl	1.018		
12346-peCl	1.239		
12347-peCl	1.259		
12348-peCl	1.230		
12349-peCl	1.331		
12367-peCl	1.214	1.226	0.012
12368-peCl	1.136		
12369-peCl	1.278		
12378-peCl	1.222	1.217	-0.005
12379-peCl	1.258		
12389-peCl	1.305		
12467-peCl	1.219		
12468-peCl	1.132	1.100	-0.032
12469-peCl	1.215		
12478-peCl	1.186	1.164	-0.022
12479-peCl	1.228	1.181	-0.047
12489-peCl	1.257		
13467-peCl	1.184		
13468-peCl	1.109		
13469-peCl	1.250		
13478-peCl	1.172		

Table 4.9 (Cont.)

	OV-101		
Isomer	Predicted	Observed	Residual
13479-peCl	1.217		
13489-peCl	1.287		
23467-peCl	1.292	1.271	-0.021
23468-peCl	1.207		
23469-peCl	1.196		
23478-peCl	1.281	1.258	-0.023
23479-peCl	1.188		
23489-peCl	1.236	1.237	0.001
123467-heCl	1.581	1.540	-0.041
123468-heCl	1.495		
123469-heCl	1.602		
123478-heCl	1.566	1.542	-0.024
123479-heCl	1.587		
123489-heCl	1.667		
123678-heCl	1.513	1.554	0.041
123679-heCl	1.623		
123689-heCl	1.588	1.604	0.016
123789-heCl	1.694		
124678-heCl	1.517	1.453	-0.064
124679-heCl	1.550		
124689-heCl	1.520	1.494	-0.026
134678-heCl	1.474	1.454	-0.020
134679-heCl	1.580		
234678-heCl	1.606	1.603	-0.003
1234678-hpCl	1.933	1.998	0.065
1234679-hpCl	2.013		
1234689-hpCl	1.983	2.061	0.078
1234789-hpCl	2.086		

Conclusions

The retention behavior for the isomers of the polychlorinated dibenzofurans was successfully modeled for six different data sets reflecting five separate chromatographic stationary phases. The models selected to represent retention behavior were statistically valid and correlated highly with observed data. Descriptors employed were topological, electronic, and geometrical as well as the atom-based descriptors which were first used in Chapter 3 of this thesis. The statistical problem of non-linearity was addressed, analyzed and solved for the OV-101 data set. A comprehensive outlier detection scheme involving two different methods, DDG and RRA, was utilized and outliers were selectively removed from the modeling process to further enhance the predictability of the models. Predictions were generated only where appropriate and the final models all passed internal validation testing. This QSRR study provides retention data where none existed before and demonstrates the usefulness of this area of research.

References

(1) Hites, R.A. *Acct. Chem. Res.* **1990**, *23*, 194.

(2) Czuczwa, J.M.; McVeety,B.D.; Hites, R.A. *Science.* **1984**, *226*, 568-569.

(3) Buser, H.P.; Bosshardt, H.P.; Rappe, C. *Chemosphere.* **1978**, *7*, 165.

(4) Hale, M.D.; Hileman, F.D.; Mazer, T.; Shell, T.L.; Noble, R.W.; Brooks, J.J. *Anal. Chem.* **1985**, *57*, 640-648.

(5) Humppi, T.; Heinola, K. *J. Chrom.* **1985**, *331*, 410-418.

(6) Kuroki, H.; Haraguchi, K.; Masuda, Y. *Chemosphere.* **1984**, *13(4)*, 561-573.

(7) Fung, D.; Boyd, R.K.; Safe, S.; Chittim, B.G. *Bio. Mass. Spect.* **1985**, *12(6)*, 247-253.

(8) Mazer, T.; Hileman, F.D.; Noble, R.W.; Brooks, J.J. *Anal. Chem.* **1983**, *55*, 104-110.

(9) Robbat, A. Jr.; Kalogeropoulos, C. *Anal. Chem.* **1990**, *62*, 2684-2688.

(10) Rohrbaugh, R.H.; Jurs, P.C. UDRAW (QCPE Program No. 300). Indiana University, IN: Quantum Chemistry Program Exchange, 1988.

(11) Stuper, A.J.; Brugger, W.E.; Jurs, P.C. *Computer Assisted Studies of Chemical Structure and Biological Function*; Wiley Interscience: New York, 1979, 83-90.

(12) MOPAC, ver 5.0. Quantum Chemistry Program Exchange, QCPE Program No. 445.

(13) Allinger, N.L.; Yul, Y.H.; MM2/MMP2, 85-Force Field (QCPE Program No. 395). Indiana University, IN: Quantum Chemistry Program Exchange, 1985.

(14) Burkert, U; Allinger, N.L. *Molecular Mechanics*; ACS Monograph 177; American Chemical Society: Washington, DC, 1982.

(15) Kier, L.B.; Hall, L.H. *Molecular Connectivity in Structure Activity Relationships*; John Wiley & Sons, Inc: New York, 1986.

(16) Rohrbough, R.H.; Jurs, P.C. *Anal. Chem.* **1986**, *58*, 1210-1212.

(17) Rohrbaugh, R.H.; Jurs, P.C. *Anal. Chem.* **1985**, *57*, 2770-2773.

(18) Buydens, L.; Massart, D.L.; Geerlings, P. *J. Chrom. Sci.* **1985**, *23*, 304-307.

(19) Abraham, R.J.; Griffiths, L.; Loftus, P. *J. Comput. Chem.* **1982**, *3*, 407-416.

(20) Abraham, R.J.; Smith, P.E. *J. Comput. Chem.* **1988**, *9*, 288-297.

(21) Dixon, S.L.; Jurs, P.C. *Empirical Calculations of Partial Atomic Charges in Organic and Ionic Compounds*, in preparation.

(22) Stouch, T.R.; Jurs, P.C. *J. Chem. Inf. Comput. Sci.* **1986**, *26(1)*, 4.

(23) Rohrbaugh, R.H.; Jurs, P.C. *Anal. Chim. Acta.* **1987**, *199*, 99-109.

(24) Weiner, H. *J. Am. Chem. Soc.* **1947**, *69*, 17.

(25) Stanton, D.T.; Jurs, P.C. *Anal. Chem.* **1990**, *62*, 2323.

(26) Pearlman, R.S.In. *Physical Chemical Properties of Drugs*; Yalkowsky, S.H.; Sinkula, A.A.; Valvani, S.C., Eds; Marcel Dekker: New York, 1980, 321-347.

(27) Goldstein, H. *Classical Mechanics;* Addison-Wesley: Reading, MA, 1950, 144-156.

(28) Wertz, D.H.; Allinger, N.L. *Tetrahedron.* **1974**, *30*, 1579-1586.

(29) Yates, K. *Hückel Molecular Orbital Theory*; Academic: New York. 1980.

(30) Lowe, J.P. *Quantum Chemistry*; Academic: New York, 1978.

(31) Neter, J.; Wasserman, W. Kutner, M.H. *Applied Linear Statistical Models*, 3rd ed; Irwin: Boston, MA, 1990.

(32) Box, G.E.P.; Cox, D.R. *J. Royal Stat. Soc.* **1964**, *B26*, 211-243.

(33) Belsey, D.A.; Kuh, E.; Welsch, R.E. *Regression DIagnostics: Identifying Influential Data and Sources of Collinearity*; Wiley Interscience: New York, 1980.

(34) Rousseeuw, P.J. *J. Am. Stat. Assoc.* **1984**, *79*, 871-880.

(35) Massart, D.L.; Kaufman, P.J.; Rouseeuw, P.J.; Lerox, A. *Anal. Chem. Acta.* **1986**, *187*, 171-179.

(36) Snee, R.D. *Technometrics.* **1977**, *4*, 415.

Chapter 5

SUMMARY

This thesis presented research in the area of quantitative structure-retention relationships. These important relationships allow for the prediction of retention behavior of various compounds on both gas and liquid chromatographic stationary phases of varying polarities. The methodology utilized for the calculations is based upon the fact that there is a definite relationship between experimentally determined retention values and the structure of the molecules. The relationships were developed with linear regression analysis which can relate the retention of a molecule on a chromatographic column to a series of descriptors which numerically encode the topology, electronics and geometry of the whole molecule. To further the process of encoding structural information, a new type of descriptor in the area of QSRR research was calculated. Atom-based descriptors were employed to describe the topological, electronic and geometrical environment of a group of selected carbon atoms only. These descriptors were able to describe the interactions of small areas of the molecule and were extremely useful in all models. These models were then used to predict the retention behavior for compounds where no experimental data existed.

In Chapter 2, Methodology, a description of the parametric approach was given. Statistical methods such as multiple linear regression analysis, objective feature selection, outlier detection and model validation techniques were discussed.

Also presented was an analysis of transformations and how they can be employed in solving statistical problems. An important aspect of this thesis was the application of the ADAPT software system and an overview of ADAPT's versatility, uniqueness, and power was provided.

In the two studies presented, regression models were developed relating retention behavior to the structural features of 210 polychlorinated dibenzodioxins and dibenzofurans. These compounds are extremely toxic and very hazardous to anyone who must handle them. Predictions of retention data instead of experimental determination is an obvious benefit to this work.

The research in Chapter 3 led to models relating the retention behavior of the 75 polychlorinated dibenzodioxins on five different stationary phases of varying polarity to structural descriptors. Models with excellent predictive ability were developed and validated for each stationary phase. The models contained topological, electronic, geometrical and atom-based descriptors. A statistical transformation was performed on one data set to enhance the validity and predictability of the model. As a result predictions of retention values were presented for isomers where experimental values were not available.

Chapter 4 also utilized the same methodology described in Chapter 2 and Chapter 3 but for the 135 polychlorinated dibenzofurans for which six data sets were available representing five different stationary phases. Excellent models were developed for all data sets. The statistical problem of non-linearity and non-constant variance was analyzed, discussed and overcome with a transformation of the dependent variable. As with Chapter 3, topological, electronic, geometrical and atom-based descriptors were employed in the final models. Predictions for all data sets were presented.

The work presented in this thesis shows the ability of QSRR and the parametric approach to chemical problem solving. Computer-based techniques are critical to this area of research. The relationships developed here with the aid of computers are only a small portion of those currently existing for other compounds. The future goal of this research is to develop adequate relationships to model retention behavior for a much broader range of compounds.